Complementary Benefits of Contrastive Learning and Self-Training Under Distribution Shift

Anonymous Authors[1]

Abstract

Self-training and contrastive learning have emerged as leading techniques for incorporating unlabeled data, both under distribution shift (unsupervised domain adaptation) and when it is absent (semi-supervised learning). However, despite the popularity and compatibility of these techniques, their efficacy in combination remains surprisingly unexplored. In this paper, we first undertake a systematic empirical investigation of this combination, finding (i) that in domain adaptation settings, self-training and contrastive learning offer significant complementary gains; and (ii) that in semi-supervised learning settings, surprisingly, the benefits are not synergistic. Across eight distribution shift datasets (*e.g.*, BREEDs, WILDS), we demonstrate that the combined method obtains 3–8% higher accuracy than either approach independently. Finally, we theoretically analyze these techniques in a simplified model of distribution shift demonstrating scenarios under which the features produced by contrastive learning can yield a good initialization for self-training to further amplify gains and achieve optimal performance, even when either method alone would fail.

1. Introduction

Even under natural, non-adversarial distribution shifts, the performance of machine learning models often drops (Quinonero-Candela et al., 2008; Torralba & Efros, 2011; Koh et al., 2021; Garg et al., 2022b). Often retraining the model on labeled data from the new distribution is impractical due to associated labeling costs. Consequently, researchers have investigated the Unsupervised Domain Adaptation (UDA) setting. Here, given labeled source data and unlabeled out-of-distribution (OOD) target data, the goal is to produce a classifier that performs well on the target. Because UDA is generally underspecified (Ben-David et al., 2010), researchers have focused on two main paths: (i) works that explore heuristics for incorporating the unlabeled target data, relying on benchmark datasets ostensibly representative of "real-world shifts" to adjudicate progress (Santurkar et al., 2021; Peng et al., 2019); and (ii) papers that explore structural assumptions under which UDA problems are well posed (Shimodaira, 2000; Schölkopf et al., 2012). This work engages with the former focusing on two popular methods: self-training and contrastive pretraining.

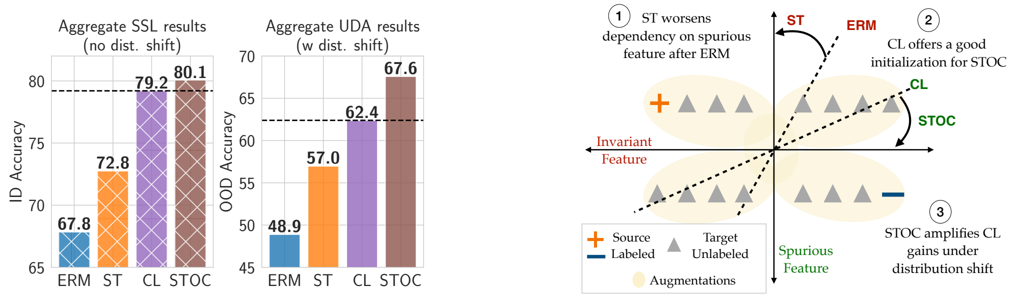
Self-training (Scudder, 1965; Lee et al., 2013; Sohn et al., 2020; Xie et al., 2020b) and contrastive pretraining (Caron et al., 2020; Chen et al., 2020a; Zbontar et al., 2021) were both proposed, initially, for traditional Semi-Supervised Learning (SSL) problems, where the labeled and unlabeled data are drawn from the same distribution. More recently, these methods have emerged as favored empirical approaches for UDA, demonstrating efficacy on many popular benchmarks (Sagawa et al., 2021; Garg et al., 2023; Cai et al., 2021; Shen et al., 2022). Several attempts have been made to understand their strong empirical performance, under various assumptions on the data, task, and inductive biases of the function class (Wei et al., 2020; HaoChen et al., 2021; Saunshi et al., 2022; Shen et al., 2022; Cai et al., 2021; HaoChen et al., 2022; HaoChen & Ma, 2022; Cabannes et al., 2023). Despite the strong results, there have been surprisingly little work (both empirically and theoretically) exploring when either might be expected to perform best and whether the benefits might be complementary.

In this paper, we investigate the complementary benefits of self-training and contrastive pretraining. Interestingly, we find that the combination yields significant gains in UDA despite producing negligible gains in SSL. In experiments across eight distribution shift benchmarks, we observe that re-using unlabeled data for self-training (with FixMatch (Sohn et al., 2020)) after learning contrastive representations (with SwAV (Caron et al., 2020)), yields $> 5\%$ average improvement on OOD accuracy in UDA as compared to $< 0.8\%$ average improvement in SSL (Fig. 1).

Next, we address the question *why the combination of self-training and contrastive learning* proves synergistic in distribution shift scenarios. To facilitate our analysis, we consider a simplified distribution shift setting that includes two

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.



(a) (b)

Figure 1: *Self-training over Contrastive learning (STOC) improves over Contrastive Learning (CL) under distribution shift.* **(a)** In SSL settings, where labeled and unlabeled data are drawn from the same distribution, STOC offers negligible gains over CL. In contrast, in UDA settings where there is a distribution shift between labeled and unlabeled data, STOC offers gains over CL. Detailed results in Table 1 and 2. **(b)** 2-D illustration of our toy distribution setup, depicting decision boundaries learned by ERM and CL and how Self-Training (ST) updates those. ①, ②, and ③ summarize our key theoretical results.

types of features: (i) invariant features that perfectly predict the label; and (ii) domain-dependent features that are predictive of the label in just source. Our theoretical analysis reveals that self-training can achieve optimal target performance but requires a "good" enough classifier to start with. We observe that source-only ERM fails to provide a "good" initialization. On the other hand, contrastive pretraining on unlabeled data performs better than ERM but is still sub-optimal. This implies that contrastive pretraining ends up decreasing reliance on domain-dependent features (as compared to ERM) but doesn't completely eliminate them. Nevertheless, contrastive pretraining does provide a "good" initialization for self-training, *i.e.*, "good" initial pseudolabels on the target unlabeled data. As a result, self-training on top of contrastive learned features effectively unlearns the reliance on domain-dependent features and generalizes perfectly OOD. In contrast, for SSL settings (*i.e.*, in distribution), our analysis highlights that contrastive pretraining already acquires sufficient predictive features such that linear probing with (a small amount of) labeled data picks up those features and attains near-optimal ID generalization.

Finally, we connect our theoretical understanding of "good" representations from contrastive learning and improved linear transferability from self-training to observed empirical gains. We linearly probe representations (fix representations and train only the linear head) learned by contrastive pretraining vs. no pretraining and find: (i) contrastive pretraining substantially improves the ceiling on the target accuracy (performance of optimal linear probe) compared to ERM; (ii) self-training mainly improves linear transfer, *i.e.* OOD accuracy of the linear probe trained with source labeled data.

1.1. Setup and Preliminaries

Task. Our goal is to learn a predictor that maps $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to $y \in \mathcal{Y}$. We parameterize predictors $f = h \circ \Phi : \mathbb{R}^d \mapsto \mathcal{Y}$, where $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^k$ is a feature map and $h \in \mathbb{R}^k$ is a classifier that maps the representation to the final scores or logits. Let $\mathrm{P_S}, \mathrm{P_T}$ be the source and target joint probability measures over $\mathcal{X} \times \mathcal{Y}$. The distribution over unlabeled samples from both the union of source and target is denoted as $\mathrm{P_U} = (1/2) \cdot \mathrm{P_S}(x) + (1/2) \cdot \mathrm{P_T}(x)$.

We study two scenarios: (i) Unsupervised Domain Adaptation (UDA); and (ii) Semi-Supervised Learning (SSL). In UDA, we assume that the source and target distributions have the same label marginals *i.e.*, $\mathrm{P_S}(y) = \mathrm{P_T}(y)$ and the same Bayes optimal predictor, *i.e.*, $\arg\max_y p_\mathsf{S}(y \mid x) = \arg\max_y p_\mathsf{T}(y \mid x)$. We are given labeled samples from the source, and unlabeled pool from the target. In SSL, there is no distribution shift, *i.e.*, $\mathrm{P_S} = \mathrm{P_T}$, and we are given a small number of labeled examples along with a comparatively large amount of unlabeled examples, both drawn from the same distribution, which we denote as $\mathrm{P_T}$. Our goal in both settings is to leverage this along with labeled data to achieve good performance on the target distribution. In the DA scenario, the challenge lies in generalizing out-of-distribution, while in SSL, the challenge is to generalize in-distribution despite the paucity of labeled examples.

Methods. We consider four algorithms (refer to App. E for precise details on the setup):

1. *Source-only ERM (ERM)*: This is standard supervised learning on labeled data by minimizing empirical risk $\sum_{i=1}^{n} \ell(h \circ \Phi(x), y)$, for some loss $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$ (*e.g.*, softmax cross-entropy) and labeled points $\{(x_i, y_i)\}_{i=1}^{n}$.

2. *Contrastive Learning (CL)*: We use unlabeled data to learn a feature extractor Φ_{cl} by optimizing an objective that maps augmentations (for e.g. crops or rotations) of the same input close to each other and far from augmentations of other random inputs (Caron et al., 2020; Chen et al., 2020a). We then learn a linear classifier h on top to minimize a classification loss on the labeled source data. We could either keep Φ_{cl} fixed or propagate gradients through. When clear from context, we also use CL to refer to just the contrastively pretrained backbone.

Table 1: *Results in the UDA setup*. We report accuracy on target (OOD) data from which we only observe unlabeled examples during training. For benchmarks with multiple target distributions (*e.g*., OH, Visda), we report avg accuracy on those targets. Refer App. F.4 for std. deviation numbers.

Method	Living17	Nonliv26	Entity13	Entity30	FMoW (2 tgts)	Visda (2 tgts)	OH (3 tgts)	CIFAR→ CINIC	Avg
ERM	60.31	45.54	68.32	55.75	56.50	20.91	9.51	74.33	48.90
ST	71.29	56.79	77.93	66.37	56.79	38.03	10.47	78.19	56.98
CL	74.14	57.02	76.58	66.01	61.78	63.49	22.63	77.51	62.39
STOC (ours)	**82.22**	**62.23**	**81.84**	**72.00**	**65.25**	**70.08**	**27.12**	**79.94**	**67.59**

3. *Self-training (ST)*: This is a two-stage procedure, where first stage does source-only ERM using source-labeled data. In the second stage, we iteratively apply the current classifier on the unlabeled data to generate "pseudolabels" and then update the classifier by minimizing a classification loss on the pseudolabeled data.

4. *Self-Training Over Contrastive learning (STOC)*: Finally, rather than starting with a source-only ERM classifier, we propose to initialize ST with CL classifier that was pretrained on unlabeled data from source and target. Now, ST uses target unlabeled data again for pseudolabeling.

2. Self-Training Improves Contrastive Pretraining Under Distribution Shift

Datasets. We conduct experiments across eight benchmark datasets: four BREEDs datasets (Santurkar et al., 2021)—Entity13, Entity30, Nonliving26, Living17; FMoW (Koh et al., 2021; Christie et al., 2018); Officehome (Venkateswara et al., 2017); Visda (Peng et al., 2018; 2017); and CIFAR-10 (Krizhevsky & Hinton, 2009). Each of these datasets consists of several domains, enabling us to construct source-target pairs (e.g., CIFAR10, we consider CIFAR10→CINIC shift (Darlow et al., 2018)). More details about datasets are in App. F.2. Because the SSL setting lacks distribution shift, we default to using source alone. To simulate limited supervision in SSL, we sub-sample the original labeled training set to 10%.

Experimental Setup and Protocols. SwAV (Caron et al., 2020) is the specific algorithm that we use for contrastive pretraining. In all UDA settings, unless otherwise specified, we pool all the (unlabeled) data from the source and target to perform SwAV. For self-training, we apply FixMatch (Sohn et al., 2020). For SSL settings, we perform SwAV and FixMatch on in-distribution unlabeled data. We experiment with Resnet18, Resnet50 (He et al., 2016) trained from scratch (*i.e*. random initialization). Moreover, unless otherwise specified, we default to full finetuning with source-only ERM, both from scratch and after contrastive pretraining, and for ST with FixMatch. For more details on model architectures, and experimental protocols, see App. F.

Results on UDA setup. Both ST and CL individually improve over ERM across all datasets, with CL significantly performing better than ST on 5 out of 8 benchmarks (see Table 1). Even on datasets where ST is better than CL, their performance remains close. Combining ST and CL with STOC shows 3–8% improvement over the best alternative, yielding improvement of 5.2% in average accuracy. In App. F.4, we highlight the significance of unlabeled target data in contrastive pretraining, where we experiment with CL model trained solely on unlabeled source data.

Table 2: *Results in the SSL setup*. We report accuracy on hold-out ID data. Recall that SSL uses labeled and unlabeled data from the same distribution during training. Refer to App. F.5 for ERM and ST with std. deviation numbers.

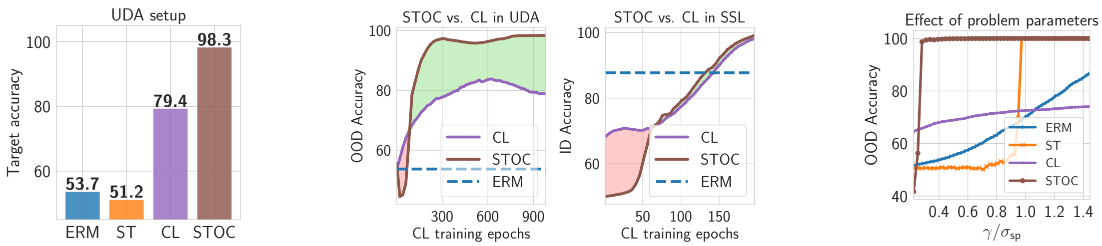
Method	Living17	Nonliv26	Entity13	Entity30	FMoW	Visda	OH	CIFAR	Avg
CL	91.15	84.58	90.73	85.47	43.05	97.67	49.73	91.78	79.27
STOC (ours)	92.00	85.95	91.27	86.14	44.43	97.70	49.95	93.06	80.06

Results on SSL setup. While CL improves over ST (as in UDA), unlike UDA, STOC doesn't offer any significant improvements over CL (see Table 2); ERM and ST results (refer to App. F.5). We conduct ablation studies with varying proportions of labeled data used for SSL, illustrating that there's considerable potential for improvement. These findings highlight that the complementary nature of STOC over CL and ST individually is an artifact of distribution shift.

3. Theoretical Analysis and Intuitions

Data distribution. We consider binary classification and model inputs as: $x = [x_{\text{in}}, x_{\text{sp}}]$, where $x_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$ is the invariant feature that is predictive of label y on both source $\mathrm{P_S}$ and target $\mathrm{P_T}$ and $x_{\text{sp}} \in \mathbb{R}^{d_{\text{sp}}}$ is the spurious feature that is only correlated with y on source. Formally, we sample $\mathrm{y} \sim \mathrm{Unif}\{-1, 1\}$ and generate x in source as $\mathrm{P_S}: x_{\text{in}} \sim \mathcal{N}(\gamma \cdot \mathrm{y} w^\star, \Sigma_{\text{in}})$, $x_{\text{sp}} = \mathrm{y}\mathbf{1}_{d_{\text{sp}}}$ and in target as $\mathrm{P_T}$: $x_{\text{in}} \sim \mathcal{N}(\gamma \cdot \mathrm{y} w^\star, \Sigma_{\text{in}})$, $x_{\text{sp}} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\text{sp}})$. Here, γ is the margin afforded by the invariant feature whose covariance is $\Sigma_{\text{in}} = \sigma_{\text{in}}^2 \cdot (\mathbf{I}_{d_{\text{in}}} - w^\star w^{\star\top})$. The spurious feature is distributed as Gaussian in the target data with $\Sigma_{\text{sp}} = \sigma_{\text{sp}}^2 \mathbf{I}_{d_{\text{sp}}}$. For convenience, we assume access to infinite unlabeled data. For SSL, we additionally sample finite labeled from $\mathrm{P_T}$ where spurious features are absent and for UDA, we assume access to infinite labeled data from the source.

Methods. We consider linear feature extractor, *i.e*. $\Phi \in \mathbb{R}^{d \times k}$, linear layer $h : \mathbb{R}^k \to \mathbb{R}$ over it, and the prediction as $\mathrm{sgn}(h^\top \Phi x)$. We use the exponential loss $\ell(f(x), y) = \exp(-yf(x))$. For ERM and ST, we train both h and Φ (equivalent to Φ being identity and training a linear head). We obtain $\Phi_{\text{cl}} := \arg\min_\Phi \mathcal{L}_{\text{cl}}(\Phi)$ by minimizing the Barlow Twins objective (Zbontar et al., 2021). The augmentation distribution $\mathrm{P_A}(a \mid x)$ scales the magnitude of each co-ordinate of x uniformly by an independent amount, i.e., $a \sim \mathrm{P_A}(\cdot \mid x) = \mathbf{c} \odot x$, where $\mathbf{c} \sim \mathrm{Unif}[0, 1]^d$. We try to mirror practical settings where the augmentations are fairly "generic". Keeping the Φ_{cl} fixed, we then learn a linear clas-



(a) (b) (c)

Figure 2: *Our simplified model of shift captures real-world trends and theoretical behaviors:* **(a)** Target (OOD) accuracy separation in the UDA setup (for problem parameters in Example G.1). **(b)** Comparison of the benefits of STOC (ST over CL) over just CL in UDA and SSL settings, done across training iterations for contrastive pretraining. **(c)** Comparison between different methods in UDA setting, as we vary problem parameters γ and σ_{sp}, connecting our theory results in Sec. 3.

sifier h_{cl} over Φ_{cl} to minimize the exponential loss on labeled source data (refer to as *linear probing*). For STOC, keeping the Φ_{cl} fixed and initializing the linear head with the CL linear probe (instead of source only ERM), we perform ST. For precise details on the objectives used for each method, along with problem parameters chosen for the data distribution see App. G.1.

3.1. Simulations and Intuitive Story

Our setup captures real-world trends in the UDA setting (see Fig. 2(a)). Before we present intuitions for this, we discuss ablating over $\gamma/\sigma_{\text{sp}}$ which is higher for easier problems.

Effect of $\gamma/\sigma_{\text{sp}}$ on success of ST. By increasing the ratio of margin γ and variance of spurious feature on target σ_{sp} (keeping others constant), the problem becomes easier because γ directly affects the signal on x_{in} and reducing σ_{sp} helps ST to unlearn x_{sp} (see App. G.3). In Fig. 2(c), we see that a phase transition occurs for ST, *i.e.*, after a certain threshold of $\gamma/\sigma_{\text{sp}}$, ST successfully recovers the optimal target predictor. This hints that ST has a binary effect, where beyond a certain magnitude of $\gamma/\sigma_{\text{sp}}$, ST can amplify the signal on domain invariant feature to obtain optimal target predictor. On the other hand, the performance of CL and ERM improve gradually where CL achieves high performance even at small ratios of $\gamma/\sigma_{\text{sp}}$. One way of viewing this trend with CL is that it magnifies the effective $\gamma/\sigma_{\text{sp}}$ in its representation space, because of which a linear head trained these representations have a good performance at low values of the ratio. Consequently, the *phase transition* of STOC occurs much sooner then that of ST. Finally, we note that for CL the rate of performance increase diminishes at high values of $\gamma/\sigma_{\text{sp}}$ because CL fails to reduce dependency along x_{sp} beyond a certain point.

An intuitive story. We return to the question of why self-training improves over contrastive learning under distribution shift in Example G.1. When the classifier at initialization of ST relies more on spurious features, ST aggravates this dependency. However, as the problem becomes easier (with increasing $\gamma/\sigma_{\text{sp}}$), the source-only ERM classifier will start relying more on invariant rather than spurious feature. Once this ERM classifier is sufficiently accurate on the target, ST unlearns any dependency on spurious features achieving optimal target accuracy. In contrast, we observe that CL performs better than ERM but is still sub-optimal. This implies that CL ends up decreasing reliance on spurious features (as compared to ERM) but doesn't completely eliminate them. Combining ST and CL, a natural hypothesis explaining our trends is that CL provides a "favorable" initialization for ST by sufficiently increasing signal on invariant features.

Why disparate behaviors for out-of-distribution vs. in distribution? In the SSL setup, recall, there is no distribution shift. In Example G.1, we sample $50k$ unlabeled data and 100 labeled data from the same (target) distribution to simulate SSL setup. Substantiating our findings on real-world data, we observe that STOC provides a small to negligible gain over CL (refer to App. G). To understand why such disparate behaviors emerge, recall that in the UDA setting, the main benefit of STOC lies in picking up reliance on "good" features for OOD data, facilitated by CL initialization. While contrastive pretraining uncovers features that are "good" for OOD data, it also learns more predictive source-only features (which are not predictive at all on target). As a result, linear probing with source-labeled data picks up these source-only features, leaving considerable room for improvement on OOD data with further self-training. On the other hand, in the SSL setting, the limited ID labeled data might provide enough signal to pick up features predictive on ID data, leaving little to no room for improvement for further self-training. Corroborating our intuitions, throughout the CL training in the toy setup, when CL doesn't achieve near-perfect generalization, the improvements provided by STOC for each checkpoint remain minimal. Contrary, for UDA setup, after reaching a certain training checkpoint in CL, STOC yields significant gains (Fig. 2(b)).

In App. G.3, G.4 we provide more results and in App. H, we formally analyze why ST and CL offer complementary benefits when dealing with distribution shifts.

References

Alexandari, A., Kundaje, A., and Shrikumar, A. Adapting to label shift with bias-corrected calibration. In *International Conference on Machine Learning (ICML)*, 2021.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Bekker, J. and Davis, J. Learning from positive and unlabeled data: a survey. *Machine Learning*, 2020. URL https://doi.org/10.1007%2Fs10994-020-05877-5.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility Theorems for Domain Adaptation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Cabannes, V., Kiani, B. T., Balestriero, R., LeCun, Y., and Bietti, A. The ssl interplay: Augmentations, inductive bias, and generalization. *arXiv preprint arXiv:2302.02774*, 2023.

Cai, T., Gao, R., Lee, J., and Lei, Q. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pp. 1170–1182. PMLR, 2021.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chapelle, O., Scholkopf, B., and Zien, A. Semi-supervised learning. 2006. *Cambridge, Massachusettes: The MIT Press View Article*, 2, 2006.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, X., Chen, W., Chen, T., Yuan, Y., Gong, C., Chen, K., and Wang, Z. Self-pu: Self boosted and calibrated positive-unlabeled training. In *International Conference on Machine Learning*, pp. 1510–1519. PMLR, 2020b.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519, 2014.

Cortes, C., Mansour, Y., and Mohri, M. Learning Bounds for Importance Weighting. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Deledalle, C.-A., Denis, L., Tabti, S., and Tupin, F. *Closed-form expressions of the eigen decomposition of 2 x 2 and 3 x 3 Hermitian matrices*. PhD thesis, Université de Lyon, 2017.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *International Conference Knowledge Discovery and Data Mining (KDD)*, pp. 213–220, 2008.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 2016.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Garg, S., Wu, Y., Smola, A., Balakrishnan, S., and Lipton, Z. Mixture proportion estimation and PU learning: A modern approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Garg, S., Balakrishnan, S., and Lipton, Z. Domain adaptation under open set label shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Garg, S., Balakrishnan, S., Lipton, Z., Neyshabur, B., and Sedghi, H. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022b.

Garg, S., Erickson, N., Sharpnack, J., Smola, A., Balakrishnan, S., and Lipton, Z. Rlsbench: A large-scale empirical study of domain adaptation under relaxed label shift. In *International Conference on Machine Learning (ICML)*, 2023.

Garrido, Q., Chen, Y., Bardes, A., Najman, L., and Lecun, Y. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022.

Grandvalet, Y. and Bengio, Y. Entropy regularization., 2006.

Gretton, A., Smola, A. J., Huang, J., Schmittfull, M., Borgwardt, K. M., and Schölkopf, B. Covariate Shift by Kernel Mean Matching. *Journal of Machine Learning Research (JMLR)*, 2009.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

HaoChen, J. Z. and Ma, T. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Joachims, T. et al. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pp. 200–209, 1999.

Johnson, D. D., Hanchi, A. E., and Maddison, C. J. Contrastive learning can find an optimal basis for approximately view-invariant functions. *arXiv preprint arXiv:2210.01883*, 2022.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Krizhevsky, A. and Hinton, G. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009.

Kschischang, F. R. The complementary error function. *Online, April*, 2017.

Kumar, A., Ma, T., and Liang, P. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Lipton, Z. C., Wang, Y.-X., and Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2017.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ma, M. Q., Tsai, Y.-H. H., Liang, P. P., Zhao, H., Zhang, K., Salakhutdinov, R., and Morency, L.-P. Conditional contrastive learning for improving fairness in self-supervised learning. *arXiv preprint arXiv:2106.02866*, 2021.

Mishra, S., Saenko, K., and Saligrama, V. Surprisingly simple semi-supervised domain adaptation with pretraining and consistency. *arXiv preprint arXiv:2101.12727*, 2021.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge, 2017.

Peng, X., Usman, B., Saito, K., Kaushik, N., Hoffman, J., and Saenko, K. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. Mit Press, 2008.

Roberts, M., Mani, P., Garg, S., and Lipton, Z. Unsupervised learning under latent label shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rosenfeld, E., Ravikumar, P., and Risteski, A. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 2002.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., Beery, S., David, E., Stavness, I., Guo, W., Leskovec, J., Saenko, K., Hashimoto, T., Levine, S., Finn, C., and Liang, P. Extending the wilds benchmark for unsupervised adaptation. In *NeurIPS Workshop on Distribution Shifts*, 2021.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations (ICLR)*, 2021.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On Causal and Anticausal Learning. In *International Conference on Machine Learning (ICML)*, 2012.

Scudder, H. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Shen, K., Jones, R. M., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 19847–19878. PMLR, 2022.

Shimodaira, H. Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, 2000.

Shu, R., Bui, H. H., Narui, H., and Ermon, S. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020.

Stewart, G. W. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 2016.

Sun, B., Feng, J., and Saenko, K. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*. Springer, 2017.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Van Engelen, J. E. and Hoos, H. H. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020a.

Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., and Zheng, Y. Self-supervised cyclegan for object-preserving image-to-image domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pp. 498–513. Springer, 2020b.

Yang, X., Song, Z., King, I., and Xu, Z. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Zadrozny, B. Learning and Evaluating Classifiers Under Sample Selection Bias. In *International Conference on Machine Learning (ICML)*, 2004.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Zhang, J., Menon, A., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations (ICLR)*, 2021.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain Adaptation Under Target and Conditional Shift. In *International Conference on Machine Learning (ICML)*, 2013.

Zhang, R. Making convolutional networks shift-invariant again. In *ICML*, 2019.

Zhang, W., Ouyang, W., Li, W., and Xu, D. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*. PMLR, 2019.

Zhu, X. and Ghahramani, Z. Learning from labeled and unlabeled data with label propagation. *CMU CALD tech report CMU-CALD-02-107, 2002*, 07 2003.

Appendix

Appendix Outline

A. Broader Impacts and Limitations of Our work

In this study, we highlight the synergistic behavior of self-training and contrastive pretraining under distribution shift. Shifts in distribution are commonplace in real-world applications of machine learning, and even under natural, non-adversarial distribution shifts, the performance of machine learning models often drops. By simply combining existing techniques in self-training and constrastive learning, we find that we can improve accuracy by 3–8% rather than using either approach independently. Despite these significant improvements, we note that one limitation of this combined approach is that performing self-training sequentially after contrastive pretraining increases the computation cost for UDA. The potential for integrating these benefits into one unified training paradigm is yet unclear, presenting an interesting direction for future exploration.

Beyond this, we note that our theoretical framework primarily confines the analysis to training the backbone and linear network independently during the pretraining and fine-tuning/self-training phases. Although our empirical observations apply to deep networks with full fine-tuning, we leave a more rigorous theoretical study of full fine-tuning for future work. Our theory also relies on a covariate shift assumption (where we assume that label distribution also doesn't shift). Investigating the complementary nature of self-training and contrastive pretraining beyond the covariate shift assumption would be another interesting direction for future work.

B. Connecting Experimental Gains with Theoretical Insights

Our theory emphasizes that under distribution shift contrastive pretraining improves the representations for target data, while self-training primarily improves linear classifiers learned on top. To investigate different methods in our UDA setup, we study the representations learned by each of them. We fix the representations and train linear heads over them to answer two questions: (i) How good are the representations in terms of their *ceiling* of target accuracy (performance of the optimal linear probe)?—we evaluate this by training the classifier head on target labeled data (*i.e.*, target linear probe); and (ii) How well do heads trained on source generalize to target?—we assess this by training a head on source labeled data (source linear probe) and evaluate its difference with target linear probe. For both, we plot target accuracy. We make two *intriguing* observations Fig. 3):

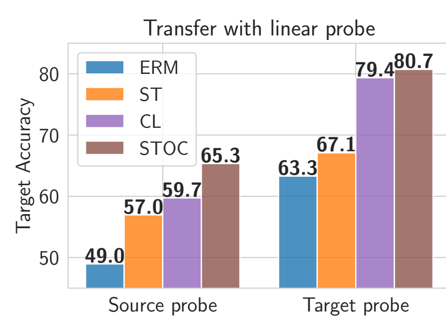


Figure 3: *Target accuracy with source and target linear probes*, which freezes backbones trained with various objectives and trains only the head in UDA setup. Avg. accuracy across all datasets. We observe that: (i) ST improves the linear transferability of source probes, and (ii) CL improves representations.

Does CL improve representations over ERM features? Yes. We observe a substantial difference in accuracy ($\approx 14\%$ gap) of target linear probes on backbones trained with contrastive pretraining (*i.e*. CL, STOC) and without it (*i.e*., ERM, ST) highlighting that CL significantly pushes the performance ceiling over non-contrastive features. As a side, our findings also stand in contrast to recent studies suggesting that ERM features might be "good enough" for OOD generalization (Rosenfeld et al., 2022; Kirichenko et al., 2022). Instead, the observed gains with contrastively pretrained backbones (*i.e*. CL, STOC) demonstrate that target unlabeled data can be leveraged to further improve over ERM features.

Do CL features yield *perfect* linear transferability from source to target? Recent works (HaoChen et al., 2022; Shen et al., 2022) conjecture that under certain conditions CL representations, linear probes learned with source labeled data may transfer perfectly from source to target. However, we observe that this doesn't hold strictly in practice, and in fact, the linear transferability can be further improved with ST. We first note a significant gap between the performance of source linear probes and target linear probes illustrating that linear transferability is not perfect in practice. Moreover, while the accuracy of target linear probes doesn't change substantially between CL and STOC, the accuracy of the source linear probe improves significantly. Similar observations hold for ERM and ST, methods trained without contrastive pretraining. This highlights that ST performs "feature refinement" to improve source to target linear transfer (with relatively small improvements in their respective target probe performance). *The findings highlight the complementary nature of benefits on real-world data: ST improves linear transferability while CL improves representations.*

C. Connections to Prior Analysis

Prior works (HaoChen et al., 2022; Shen et al., 2022) analyzing CL first make assumptions on the consistency of augmentations with labels (HaoChen et al., 2021; Cabannes et al., 2023; Saunshi et al., 2022; Johnson et al., 2022), and specifically for UDA make stronger ones on the augmentation graph connecting examples from same domain or class more than cross-class/cross-domain ones. While this is sufficient to prove linear transferability, it is unclear if this holds in practice when augmentations are imperfect, *i.e*. if they fail to mask the spurious features completely—as corroborated by our findings in Sec. B. We show why this also fails in our simplified setup in App. I.1. Some prior works on self-training view it as consistency regularization that constrains pseudolabels of original samples to be consistent with all their augmentations (Cai et al., 2021; Wei et al., 2020; Sohn et al., 2020). Since this framework does not account challenges of propagating labels (*e.g*., when augmentation distribution has long tails) iteratively for deep networks, in our analysis we instead adopt the iterative analysis of self-training (Chen et al., 2020b) (for more discussion see App. I.2). Notably, our empirical results and our analyses offer a perspective that contrasts with the prior literature that argue for the individual optimality of contrastive pretraining and self-training. We expand on this and other related works in App. D.

D. Other Related Works

Unsupervised domain adaption. Without assumption on the nature of shift, UDA is underspecified (Ben-David et al., 2010). This challenge has been addressed in various ways by researchers. One approach is to investigate additional structural assumptions under which UDA problems are well posed (Shimodaira, 2000; Schölkopf et al., 2012). Popular settings for which DA is well-posed include (i) *covariate shift* (Zhang et al., 2013; Zadrozny, 2004; Cortes et al., 2010; Cortes & Mohri, 2014; Gretton et al., 2009) where $p(x)$ can change from source to target but $p(y|x)$ remains invariant; and (ii) *label shift* (Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2021; Garg et al., 2020; Zhang et al., 2021; Roberts et al., 2022; Garg et al., 2023) where the label marginal $p(y)$ can change but $p(x|y)$ is shared across source and target. Principled methods with strong theoretical guarantees exists for adaptation under these settings when target distribution's support is a subset of the source support. Other works (Elkan & Noto, 2008; Bekker & Davis, 2020; Garg et al., 2021; 2022a) extend the label shift setting to scenarios where previously unseen classes may appear in the target and $p(x|y)$ remains invariant among seen classes. A complementary line of research focuses on constructing benchmarks to develop heuristics for incorporating the unlabeled target data, relying on benchmark datasets ostensibly representative of "real-world shifts" to adjudicate progress (Santurkar et al., 2021; Venkateswara et al., 2017; Sagawa et al., 2021; Peng et al., 2019; 2017). As a result, various benchmark-driven heuristics have been proposed (Long et al., 2015; 2017; Sun & Saenko, 2016; Sun et al., 2017; Zhang et al., 2019; 2018; Ganin et al., 2016; Sohn et al., 2020). Our work engages with the latter, focusing on two popular methods: self-training and contrastive pretraining.

Domain generalization. In domain generalization, the model is given access to data from multiple different domains and the goal is to generalize to a previously unseen domain at test time (Blanchard et al., 2011; Muandet et al., 2013). For a survey of different algorithms for domain generalization, we refer the reader to Gulrajani & Lopez-Paz (2020). A crucial distinction here is that unlike the domain generalization setting, in DA problems, we have access to unlabeled examples from the test domain.

Semi-supervised learning. To learn from a small amount of labeled supervision, semi-supervised learning methods leverage unlabeled data alongside to improve learning models. One of the seminal works in SSL is the pseudolabeling method (Scudder, 1965), where a classifier is trained on the labeled data and then used to classify the unlabeled data, which are then added to the training set. The work of Zhu & Ghahramani (2003) built on this by introducing graph-based methods, and the transductive SVMs (Joachims et al., 1999) presented an SVM-based approach. More recent works have focused on deep learning techniques, and similar to UDA, self-training and contrastive pretraining have emerged as two prominent choices. We delve into these methods in greater detail in the following paragraphs. For a discussion on other SSL methods, we refer interested readers to (Chapelle et al., 2006; Van Engelen & Hoos, 2020; Yang et al., 2022).

Self-training. Two popular forms of self-training are pseudolabeling (Lee et al., 2013) and conditional entropy minimization (Grandvalet & Bengio, 2006), which have been observed to be closely connected (Berthelot et al., 2019; Lee et al., 2013; Sohn et al., 2020; Shu et al., 2018). Motivated by its strong performance in SSL and UDA settings (Sohn et al., 2020; Xie et al., 2020a; Garg et al., 2023; Shu et al., 2018), several theoretical works have made attempts to understand its behavior (Kumar et al., 2020; Wei et al., 2020; Chen et al., 2020b). (Wei et al., 2020; Cai et al., 2021) aims to understand the behavior of the global minimizer of self-training objective by studying input consistency regularization, which enforces stability of the prediction for different augmentations of the unlabeled data. Our analysis of self-training is motivated by the work of Chen et al. (2020b) which explores the iterative behavior of self-training to unlearn spurious features. The setting of spurious features is of particular interest, since prior works have specifically analyzed the failures of out-of-distribution generalization in the presence of spurious features (Nagarajan et al., 2020; Sagawa et al., 2020).

Contrastive learning. An alternate line of work that uses unlabeled data for learning representations in the pretraining stage is contrastive learning (Grill et al., 2020; Oord et al., 2018; Caron et al., 2020; Chen et al., 2020a; Wu et al., 2018). Given an augmentation distribution, the main goal of contrastive objectives is to map augmentations drawn from the same input (positive pairs) to similar features, and force apart features corresponding to augmentations of different inputs (negative pairs) (Caron et al., 2020; 2021; He et al., 2020). Prior works (Cabannes et al., 2023; Johnson et al., 2022; HaoChen & Ma, 2022) have also shown a close relationship between contrastive (Chen et al., 2020a; HaoChen et al., 2021) and non-contrastive objectives (Bardes et al., 2021; Zbontar et al., 2021). Consequently, in our analysis pertaining to the toy setup we focus on the mathematically non-contrastive objective Barlow Twins (Zbontar et al., 2021). Using this pretrained backbone (either as an initialization or as a fixed feature extractor) a downstream predictor is learned using labeled examples. Several works (HaoChen et al., 2021; Saunshi et al., 2022; HaoChen & Ma, 2022; Arora et al., 2019; Johnson et al., 2022) have analyzed the in-distribution generalization of the downstream predictor via label consistency arguments on the graph of positive pairs (augmentation graph). In contrast, we study the impact of contrastive learning under distribution shifts in the UDA setup. Other works (Shen et al., 2022; HaoChen et al., 2022) that examine contrastive learning for UDA also conform to the augmentation graph view point, making additional assumptions that guarantee linear transferability. In our simplified setup involving spurious correlations, these abstract assumptions break easily when the augmentations are of a generic nature, akin to practice. Finally, some empirical works (Mishra et al., 2021; Ma et al., 2021) have found self-supervised objectives like contrastive pretraining to reduce dependence on spurious correlations. Corroborating their findings, we extensively evaluate the complementary benefits of contrastive learning and self-training on real-world datasets. Finding differing results in SSL and UDA settings, we further examine their behavior theoretically in our toy setup.

E. More Details on Problem Setup

In this section, we elaborate on our setup and methods studied in our work.

Task. Our goal is to learn a predictor that maps inputs $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to outputs $y \in \mathcal{Y}$. We parameterize predictors $f = h \circ \Phi : \mathbb{R}^d \mapsto \mathcal{Y}$, where $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^k$ is a feature map and $h \in \mathbb{R}^k$ is a classifier that maps the representation to the final scores or logits. Let $\mathrm{P_S}, \mathrm{P_T}$ be the source and target joint probability measures over $\mathcal{X} \times \mathcal{Y}$ with p_S and p_T as the corresponding probability density (or mass) functions. The distribution over unlabeled samples from both the union of source and target is denoted as $\mathrm{P_U} = (1/2) \cdot \mathrm{P_S}(x) + (1/2) \cdot \mathrm{P_T}(x)$.

We study two particular scenarios:

Unsupervised Domain Adaptation (UDA). We assume that we are given labeled data from the *source* distribution and unlabeled data from a shifted, *target* distribution, with the goal of performing well on target data. We assume that the source and target distributions have the same label marginals $\mathrm{P}_{\mathsf{S}}(y) = \mathrm{P}_{\mathsf{T}}(y)$ (*i.e.*, no label proportion shift) and the same Bayes optimal predictor, *i.e.*, $\arg\max_y p_{\mathsf{S}}(y \mid x) = \arg\max_y p_{\mathsf{T}}(y \mid x)$. We are given labeled samples from the source, and unlabeled pool from the target. Here, even with infinite labeled source data, the challenge lies in generalizing out-of-distribution. In experiments, we assume access to finite data but in theory, we assume population access to labeled source and unlabeled target.

Semi-Supervised Learning (SSL). Here, there is no distribution shift, *i.e.*, $\mathrm{P}_{\mathsf{S}} = \mathrm{P}_{\mathsf{T}}$. We are given a small number of labeled examples and a comparatively large amount of unlabeled examples, both drawn from the same distribution. Without loss of generality, we denote this distribution with P_{T}. The goal in SSL is to generalize in-distribution. The challenge is primarily due to limited access to labeled data. Here, in experiments, we assume limited access to labeled data but a comparatively larger amount of unlabeled in-distribution data. In theory, we assume population access to unlabeled data but limited labeled examples.

Unlabeled data is typically much cheaper to obtain, and our goal in both these settings is to leverage this along with labeled data to achieve good performance on the target distribution. In the DA scenario, the challenge lies in generalizing out-of-distribution, while in SSL, the challenge is to generalize in-distribution despite the paucity of labeled examples. A predictor f is evaluated on distribution P via its accuracy, *i.e.*, $A(f, \mathrm{P}) = \mathbb{E}_{\mathrm{P}}(\arg\max f(x) = y)$.

Methods. We now introduce the algorithms used for learning from labeled and unlabeled data.

1. *Source-only ERM (ERM)*: A standard approach is to simply perform supervised learning on the labeled data by minimizing the empirical risk $\sum_{i=1}^{n} \ell(h \circ \Phi(x), y)$, for some classification loss $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$ (*e.g.*, softmax cross-entropy) and labeled points $\{(x_i, y_i)\}_{i=1}^{n}$.
2. *Contrastive Learning (CL)*: We first use the unlabeled data to learn a feature extractor. In particular, the objective is to learn a feature extractor Φ_{cl} that maps augmentations (for e.g. crops or rotations) of the same input close to each other and far from augmentations of random other inputs (Caron et al., 2020; Chen et al., 2020a; Zbontar et al., 2021). Once we have Φ_{cl}, we learn a linear classifier h on top to minimize a classification loss on the labeled source data. We could either keep Φ_{cl} fixed or propagate gradients through.

   When clear from context, we also use CL to refer to just the contrastively pretrained backbone without training for downstream classification.
3. *Self-training (ST)*: This is a two-stage procedure, where the first stage performs source-only ERM by just looking at source-labeled data. In the second stage, we iteratively apply the current classifier on the unlabeled data to generate "pseudo-labels" and then update the classifier by minimizing a classification loss on the pseudolabeled data (Lee et al., 2013).
4. *Self-Training Over Contrastive learning (STOC)*: Finally, rather than starting with a source-only ERM classifier, we propose to initialize self-training with a CL classifier, that was pretrained on unlabeled source and target data. ST uses that same unlabeled data again for pseudolabeling. As we demonstrate experimentally and theoretically, this combination of methods improves substantially over each independently.

Table 8 summarizes the main methods and key differences between those methods in UDA and SSL setup. For exact implementation in our experiments, we refer reader to App. F.3.

F. Additional Experiments and Details

F.1. Additional setup and notation

Recall, our goal is to learn a predictor that maps inputs $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to outputs $y \in \mathcal{Y}$. We parameterize predictors $f = h \circ \Phi : \mathbb{R}^d \mapsto \mathcal{Y}$, where $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^k$ is a feature map and $h \in \mathbb{R}^k$ is a classifier that maps the representation to the final scores or logits. With $A : \mathcal{X} \to \mathcal{A}$, we denote the augmentation function that takes in an input x and outputs an augmented view of the input $A(x)$. Unless specified otherwise, we perform full-finetuning in all of our experiments on real-world data. That is, we backpropagate gradients in both the linear head h and the backbone ϕ. For UDA, we denote source labeled points as $\{(x_i, y_i)\}_{i=1}^{n}$ and target unlabeled points as $\{(x_i')\}_{i=1}^{m}$. For SSL, we use the same notation for labeled and unlabeled in-distribution data.

F.2. Dataset details

For both UDA and SSL, we conduct experiments across eight benchmark datasets. Each of these datasets consists of domains, enabling us to construct source-target pairs for UDA. The adopted source and target domains are standard to previous studies (Shen et al., 2022; Garg et al., 2023; Sagawa et al., 2021). Because the SSL setting lacks distribution shift, we do not need to worry about domain designations and default to using source alone. To simulate limited supervision in SSL, we sub-sample the original labeled training set to 10%. Below provide exact details about the datasets used in our benchmark study.

- **CIFAR10** We use the original CIFAR10 dataset (Krizhevsky & Hinton, 2009) as the source dataset. For target domains, we consider CINIC10 (Darlow et al., 2018) which is a subset of Imagenet restricted to CIFAR10 classes and downsampled to 32×32.
- **FMoW** In order to consider distribution shifts faced in the wild, we consider FMoW-WILDs (Koh et al., 2021; Christie et al., 2018) from WILDS benchmark, which contains satellite images taken in different geographical regions and at different times. We use the original train as source and OOD val and OOD test splits as target domains as they are collected over different time-period. Overall, we obtain 3 different domains (1 source and 2 targets).
- **BREEDs** We also consider BREEDs benchmark (Santurkar et al., 2021) in our setup to assess robustness to subpopulation shifts. BREEDs leverage class hierarchy in ImageNet (Russakovsky et al., 2015) to re-purpose original classes to be the subpopulations and defines a classification task on superclasses. We consider distribution shift due to subpopulation shift which is induced by directly making the subpopulations present in the training and test distributions disjoint. BREEDs benchmark contains 4 datasets **Entity-13**, **Entity-30**, **Living-17**, and **Non-living-26**, each focusing on different subtrees and levels in the hierarchy. Overall, for each of the 4 BREEDs datasets (i.e., Entity-13, Entity-30, Living-17, and Non-living-26), we obtain one different domain which we consider as target. We refer to source and target as follows: BREEDs sub-population 1, BREEDs sub-population 2.
- **OfficeHome** We use four domains (art, clipart, product and real) from OfficeHome dataset (Venkateswara et al., 2017). We use the product domain as source and the other domains as target.
- **Visda** We use three domains (train, val and test) from the Visda dataset (Peng et al., 2018; 2017). While 'train' domain contains synthetic renditions of the objects, 'val' and 'test' domains contain real world images. To avoid confusing, the domain names with their roles as splits, we rename them as 'synthetic', 'Real-1' and 'Real-2'. We use the synthetic (original train set) as the source domain and use the other domains as target.

We summarize the information about source and target domains in Table 3.

Dataset	Source	Target
CIFAR10	CIFAR10v1	CINIC10
FMoW	FMoW (2002–'13)	FMoW (2013–'16), FMoW (2016–'18)
Entity13	Entity13 (sub-population 1)	Entity13 (sub-population 2)
Entity30	Entity30 (sub-population 1)	Entity30 (sub-population 2),
Living17	Living17 (sub-population 1)	Living17 (sub-population 2),
Nonliving26	Nonliving26 (sub-population 1)	Nonliving26 (sub-population 2),
Officehome	Product	Product, Art, ClipArt, Real
Visda	Synthetic (originally referred to as train)	Synthetic, Real-1 (originally referred to as val), Real-2 (originally referred to as test)

Table 3: Details of source and target sets in each dataset considered in our testbed.

Train-test splits We partition each source and target dataset into 80% and 20% i.i.d. splits. We use 80% splits for training and 20% splits for evaluation (or validation). We throw away labels for the 80% target split and only use labels in the 20% target split for final evaluation. The rationale behind splitting the target data is to use a completely unseen batch of data

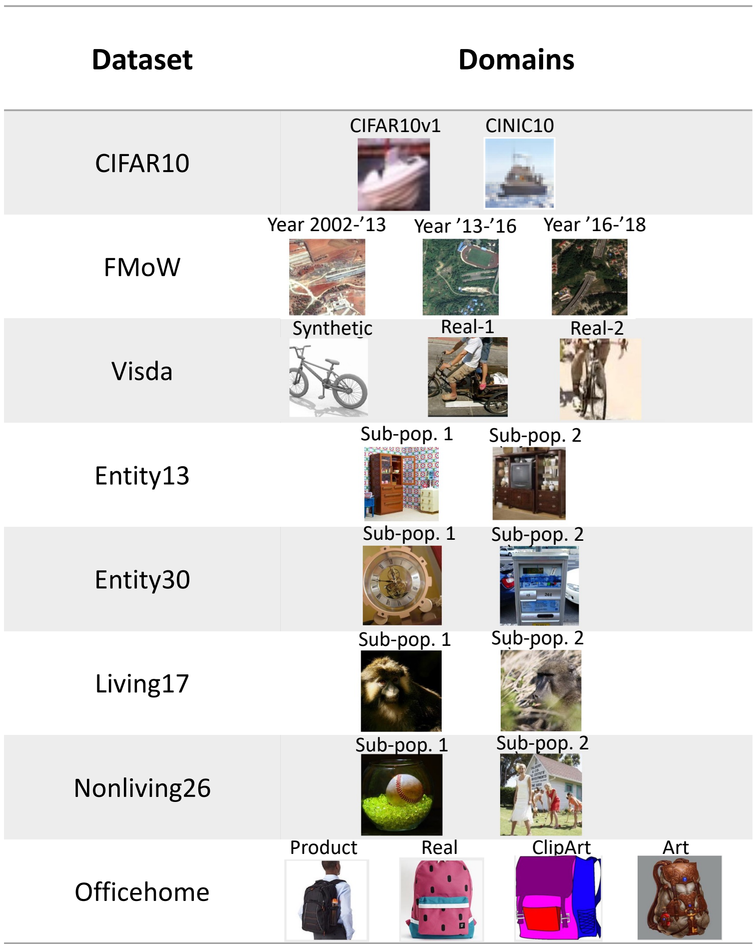


Figure 4: Examples from all the domains in each dataset.

for evaluation. This avoids evaluating on examples where a model potentially could have overfit. over-fitting to unlabeled examples for evaluation. In practice, if the aim is to make predictions on all the target data (i.e., transduction), we can simply use the (full) target set for training and evaluation.

Simulating SSL settings and limited supervision. For SSL settings, we choose the in-distribution domain as the source domain. To simulate limited supervision in SSL, we sub-sample the original labeled training set to 10% and use all the original dataset as unlabeled data. For evaluation, we further split the original holdout set into two partitions (one for validation and the other to report final accuracy numbers).

F.3. Method details

For implementation, we build on top of WILDs (Sagawa et al., 2021) and RLSbench (Garg et al., 2023) open source libraries.

ERM (Source only) training. We consider Empirical Risk Minimization (ERM) on the labeled source data as a baseline. Since this simply ignores the unlabeled target data, we call this as source only training. As mentioned in the main paper, we perform source only training with data augmentations. Formally, we minimize the following ERM loss:

$$L_{\text{source only}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(A(x_i), y_i)) \,, \tag{1}$$

where A is the stochastic data augmentation operation and ℓ is a loss function. For SSL, the ERM baseline only uses the small of labeled data available.

Contrastive Learning (CL). We perform contrastive pretraining on the unlabeled dataset to obtain the backbone ϕ_{cl}. And then we perform full fine-tuning with source labeled data by initializing the backbone with ϕ_{cl}. We use SwAV (Caron et al., 2020) for contrastive pretraining. The main idea behind SwAV is to train a model to identify different views of the same image as similar, while also ensuring that it finds different images to be distinct. This is accomplished through a *swapped* prediction mechanism, where the goal is to compute a code from an augmented version of the image and predict this code from other augmented versions of the same image. In particular, given two image features $\phi(x'_{a1})$ and $\phi(x'_{a2})$ from two different augmentations of the same image x', i.e., $x'_{a1}, x'_{a2} \sim A(x')$, SwAV computes their codes z_{a1} and z_{a2} by matching the features to a set of K prototypes $\{c_1, \cdots, c_K\}$. Then SwAV minimizes the following loss such that $\phi(x'_{a1})$ can compute codes z_{a2} and $\phi(x'_{a2})$ can compute codes z_{a1}:

$$L_{\text{SwAV}}(\phi) = \sum_{i=1}^{m} \sum_{x'_{i,a1}, x'_{i,a2} \sim A(x'_i)} \ell'(\phi(x'_{i,a1}), z_{i,a2}) + \ell'(\phi(x'_{i,a2}), z_{i,a1}) \,, \tag{2}$$

where ℓ' computes KL-divergence between codes computed with features (e.g. $\phi(x_{a1})$) and the code computed by another view (e.g. z_{a2}). For more details about the algorithm, we refer the reader to Caron et al. (2020). In all UDA settings, unless otherwise specified, we pool all the (unlabeled) data from the source and target to perform SwAV. For SSL, we leverage in-distribution unlabeled data.

We employ SimCLR (Chen et al., 2020a) for the CIFAR10 dataset, aligning with previous studies that have utilized contrastive pretraining on the same dataset (Kumar et al., 2022; Shen et al., 2022). The reason for this choice is that SwAV relies on augmentations that involve cropping images to a smaller resolution, making it more suitable for datasets with larger resolutions beyond 32×32.

Self-Training (ST). For self-training, we apply FixMatch (Sohn et al., 2020), where the loss on labeled data and on pseudolabeled unlabeled data are minimized simultaneously. Sohn et al. (2020) proposed FixMatch as a variant of the simpler Pseudo-label method (Lee et al., 2013). This algorithm dynamically generates psuedolabels and overfits on them in each batch. FixMatch employs consistency regularization on the unlabeled data. In particular, while pseudolabels are generated on a weakly augmented view of the unlabeled examples, the loss is computed with respect to predictions on a strongly augmented view. The intuition behind such an update is to encourage a model to make predictions on weakly augmented data consistent with the strongly augmented example. Moreover, FixMatch only overfits to the assigned labeled with weak augmentation if the confidence of the prediction with strong augmentation is greater than some threshold τ. Refer to A_{weak} as the weak-augmentation and A_{strong} as the strong-augmentation function. Then, FixMatch uses the following loss function:

$$\begin{aligned} L_{\text{FixMatch}}(f) = & \frac{1}{n} \sum_{i=1}^{n} \ell(f(A_{\text{strong}}(x_i), y_i)) \\ & + \frac{\lambda}{m} \sum_{i=1}^{m} \ell(f(A_{\text{strong}}(x'_i), \widetilde{y}_i)) \cdot \mathbb{I}\left[\max_y f_y(A_{\text{strong}}(x'_i)) \geqslant \tau\right] \,, \end{aligned}$$

where $\widetilde{y}_i = \arg\max_y f_y(T_{\text{weak}}(x_i))$. For UDA, our unlabeled data is the union of source and target unlabeled data. For SSL, we only leverage in-distribution unlabeled data.

We adapted our implementation from Sagawa et al. (2021) which matches the implementation of Sohn et al. (2020) except for one detail. While Sohn et al. (2020) augments labeled examples with weak augmentation, Sagawa et al. (2021) proposed to strongly augment the labeled source examples.

Self-Training Over Contrastive learning (STOC). Finally, rather than performing FixMatch from a randomly initialized backbone, we initialize FixMatch with a contrastive pretrained backbone.

F.4. Additional UDA experimemts

Table 4: *Results in the UDA setup*. We report accuracy on target (OOD) data from which we only observe unlabeled examples during training. For benchmarks with multiple target distributions (*e.g.*, OH, Visda), we report average accuracy on those targets.

Method	Living17	Nonliv26	Entity13	Entity30	FMoW (2 tgts)	Visda (2 tgts)	OH (3 tgts)	CIFAR→ CINIC
ERM	$60.2_{\pm 0.1}$	$45.4_{\pm 0.2}$	$68.6_{\pm 0.1}$	$55.7_{\pm 0.0}$	$56.5_{\pm 0.1}$	$20.8_{\pm 0.2}$	$9.5_{\pm 0.2}$	$74.3_{\pm 0.1}$
ST	$71.1_{\pm 0.2}$	$56.8_{\pm 0.1}$	$78.0_{\pm 0.3}$	$66.7_{\pm 0.1}$	$56.9_{\pm 0.4}$	$39.1_{\pm 0.1}$	$11.1_{\pm 0.1}$	$78.3_{\pm 0.3}$
CL	$74.1_{\pm 0.2}$	$57.4_{\pm 0.3}$	$76.9_{\pm 0.2}$	$66.6_{\pm 0.3}$	$61.5_{\pm 0.5}$	$63.2_{\pm 0.2}$	$22.8_{\pm 0.1}$	$77.5_{\pm 0.1}$
STOC (ours)	$\mathbf{82.6}_{\pm 0.1}$	$\mathbf{62.1}_{\pm 0.2}$	$\mathbf{81.9}_{\pm 0.2}$	$\mathbf{72.0}_{\pm 0.2}$	$\mathbf{65.3}_{\pm 0.1}$	$\mathbf{70.1}_{\pm 0.2}$	$\mathbf{27.1}_{\pm 0.3}$	$\mathbf{79.9}_{\pm 0.3}$

Note that by default, we train with CL on the combined unlabeled data from source and target. However, to better understand the significance of unlabeled target data in contrastive pretraining, we perform an ablation where the CL model was trained solely on unlabeled source data (refer to this as CL (source only); see Table 5). We observe that ST on top of CL (source only) improves over ST (from scratch). However, the average performance of ST over CL (source only) is similar to that of standalone CL, maintaining an approximate 6% performance gap observed between CL and ST. This brings two key insights to the fore: (i) the observed benefit is not merely a result of the contrastive pretraining objective alone, but specifically CL with unlabeled target data helps; and (ii) both CL and ST leverage using target unlabeled data in a complementary nature.

Table 5: *Results in the UDA setup with source only contrastive pretraining*. We report accuracy on target (OOD) data from which we only observe unlabeled examples during training. For benchmarks with multiple target distributions (*e.g.*, OH, Visda), we report average accuracy on those targets.

Method	Living17	Nonliv26	Entity13	Entity30	FMoW (2 tgts)	Visda (2 tgts)	OH (3 tgts)	CIFAR→ CINIC
CL (source only)	$67.3_{\pm 0.1}$	$49.1_{\pm 0.2}$	$71.5_{\pm 0.1}$	$58.5_{\pm 0.3}$	$53.9_{\pm 0.1}$	$33.3_{\pm 0.2}$	$21.7_{\pm 0.1}$	$77.7_{\pm 0.1}$
STOC (source only)	$75.0_{\pm 0.2}$	$58.4_{\pm 0.1}$	$79.8_{\pm 0.3}$	$67.5_{\pm 0.1}$	$56.3_{\pm 0.4}$	$42.7_{\pm 0.1}$	$25.7_{\pm 0.1}$	$77.8_{\pm 0.1}$
CL	$74.1_{\pm 0.2}$	$57.4_{\pm 0.3}$	$76.9_{\pm 0.2}$	$66.6_{\pm 0.3}$	$61.5_{\pm 0.5}$	$63.2_{\pm 0.2}$	$22.8_{\pm 0.1}$	$77.5_{\pm 0.1}$
STOC	$\mathbf{82.6}_{\pm 0.1}$	$\mathbf{62.1}_{\pm 0.2}$	$\mathbf{81.9}_{\pm 0.2}$	$\mathbf{72.0}_{\pm 0.2}$	$\mathbf{65.3}_{\pm 0.1}$	$\mathbf{70.1}_{\pm 0.2}$	$\mathbf{27.1}_{\pm 0.3}$	$\mathbf{79.9}_{\pm 0.3}$

F.5. Additional SSL experimemts

Table 6: *Results in the SSL setup*. We report accuracy on hold-out ID data. Recall that SSL uses labeled and unlabeled data from the same distribution during training.

Method	Living17	Nonliv26	Entity13	Entity30	FMoW	Visda	OH	CIFAR
ERM	$76.8_{\pm 0.1}$	$64.9_{\pm 0.2}$	$80.1_{\pm 0.0}$	$70.4_{\pm 0.3}$	$33.6_{\pm 0.4}$	$99.2_{\pm 0.0}$	$32.0_{\pm 0.2}$	$85.5_{\pm 0.1}$
ST	$85.4_{\pm 0.1}$	$75.7_{\pm 0.2}$	$85.4_{\pm 0.2}$	$77.3_{\pm 0.1}$	$33.6_{\pm 0.3}$	$99.2_{\pm 0.1}$	$32.0_{\pm 0.1}$	$93.1_{\pm 0.1}$
CL	$91.1_{\pm 0.5}$	$84.6_{\pm 0.6}$	$90.7_{\pm 0.4}$	$85.5_{\pm 0.3}$	$43.1_{\pm 0.2}$	$97.6_{\pm 0.3}$	$49.7_{\pm 0.2}$	$91.7_{\pm 0.2}$
STOC (ours)	$92.0_{\pm 0.1}$	$85.8_{\pm 0.2}$	$91.3_{\pm 0.3}$	$86.1_{\pm 0.2}$	$44.4_{\pm 0.1}$	$97.7_{\pm 0.2}$	$49.9_{\pm 0.2}$	$93.06_{\pm 0.3}$

F.6. Other experimental details

Augmentations. For weak augmentation, we leverage random horizontal flips and random crops of pre-defined size. For SwAV, we also perform multicrop augmentation as proposed in Caron et al. (2020). For strong augmentation, we apply the following transformations sequentially: random horizontal flips, random crops of pre-defined size, augmentation with Cutout (DeVries & Taylor, 2017), and RandAugment (Cubuk et al., 2020). For the exact implementation of RandAugment, we directly use the implementation of Sohn et al. (2020). Unless specified otherwise, for all methods, we default to using strong augmentation techniques.

Architectures. In our work, we experiment with Resnet18, Resnet50 (He et al., 2016) trained from scratch (*i.e.* random initialization). We do not consider off-the-shelf pretrained models (*e.g.*, on Imagenet (Russakovsky et al., 2015)) to avoid confounding our conclusions about contrastive pretraining. However, we note that our results on most datasets tend to be comparable to and sometimes exceed those obtained with ImageNet pretrained models. For BREEDs datasets, we employ Resnet18 architecture. For other datasets, we train a Resnet50 architecture.

Except for Resnets on CIFAR dataset, we used the standard pytorch implementation (Gardner et al., 2018). For Resnet on Cifar, we refer to the implementation here: `https://github.com/kuangliu/pytorch-cifar`. For all the architectures, whenever applicable, we add antialiasing (Zhang, 2019). We use the official library released with the paper.

Hyperparameters. For all the methods, we fix the algorithm-specific hyperparameters to the original recommendations. For UDA, given that the setup precludes access to labeled data from the target distribution, we use source hold-out performance to pick the best hyperparameters. During pretraining, early stopping is done according to lower values of pretraining loss.

We tune the learning rate and ℓ_2 regularization parameter by fixing the batch size for each dataset that corresponds to the maximum we can fit to 15GB GPU memory. We default to using cosine learning rate schedule (Loshchilov & Hutter, 2016). We set the number of epochs for training as per the suggestions of the authors of respective benchmarks. For SSL, we run both ERM and FixMatch for approximately 2000 epochs. Note that we define the number of epochs as a full pass over the labeled training source data. We summarize the learning rate, batch size, number of epochs, and ℓ_2 regularization parameter used in our study in Table 7.

Dataset	Batch size	ℓ_2 regularization set	Learning rate set
CIFAR10	200	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.2, 0.1, 0.05, 0.01, 0.003, 0.001\}$
FMoW	64	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.01, 0.003, 0.001, 0.0003, 0.0001\}$
Entity13	256	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005\}$
Entity30	256	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005\}$
Entity30	256	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005\}$
Nonliving26	256	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005\}$
Officehome	96	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.01, 0.003, 0.001, 0.0003, 0.0001\}$
Visda	96	$\{0.001, 0.0001, 10^{-5}, 0.0\}$	$\{0.03, 0.01, 0.003, 0.001, 0.0003\}$

Table 7: Details of the batch size, learning rate set and ℓ_2 regularization set considered in our testbed.

Compute infrastructure. Our experiments were performed across a combination of Nvidia T4, A6000, and V100 GPUs.

G. Additional Results in Toy Setup

In this section we will first give more details on our simplified setup that captures both contrastive pretraining and self-training in the same framework. Then, we provide some additional empirical results that are not captured theoretically but mimic behaviors observed in real world settings, highlighting the richness of our setup.

G.1. Detailed description of our simplified setup

In this subsection, we will first re-iterate the problem setup in Sec. 3 and provide some comparisons between our setup and those in closely related works. We will then describe the four methods: ERM, ST, CL, and STOC, providing details on the exact estimates returned by these algorithms in the SSL and UDA settings.

Our results on real-world datasets suggest that although self-training may offer little to no improvement over contrastive pretraining for in-distribution (*i.e.*, SSL) settings, it leads to substantial improvements when facing distribution shifts in UDA (Sec. 2). Why do these methods offer complementary gains, but only under distribution shifts? In this section, we seek to answer this question by first replicating all the empirical trends of interest in a simple data distribution with an intuitive story (Sec. 3.1). In this toy model, we formally characterize the gains afforded by contrastive pretraining and self-training both individually (Secs. H.1, H.2) and when used together (Sec. H.3).

Data distribution. We consider binary classification and model the inputs as consisting of two kinds of features: $x = [x_{\text{in}}, x_{\text{sp}}]$ where $x_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$ is the invariant feature that is predictive of the label across both source $\mathrm{P_S}$ and target $\mathrm{P_T}$ and $x_{\text{sp}} \in \mathbb{R}^{d_{\text{sp}}}$ is the spurious feature that is correlated with the label y only on the source domain $\mathrm{P_S}$ but uncorrelated with label y in $\mathrm{P_T}$. Here, $x_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$ determines the label using the ground truth classifier $w^\star \sim \mathrm{Unif}(\mathbb{S}^{d_{\text{in}}-1})$, and $x_{\text{sp}} \in \mathbb{R}^{d_{\text{sp}}}$ is strongly correlated with the label on source but random noise on target. Formally, we sample $\mathrm{y} \sim \mathrm{Unif}\{-1, 1\}$ and generate inputs x conditioned on y as follows

$$\mathrm{P_S}: \quad x_{\text{in}} \sim \mathcal{N}(\gamma \cdot \mathrm{y} w^\star, \Sigma_{\text{in}}) \quad x_{\text{sp}} = \mathrm{y}\mathbf{1}_{d_{\text{sp}}} \tag{3}$$

$$\mathrm{P_T}: \quad x_{\text{in}} \sim \mathcal{N}(\gamma \cdot \mathrm{y} w^\star, \Sigma_{\text{in}}) \quad x_{\text{sp}} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\text{sp}}), \tag{4}$$

where γ is the margin afforded by the invariant feature. We set covariance of the invariant features $\Sigma_{\text{in}} = \sigma_{\text{in}}^2 \cdot (\mathbf{I}_{d_{\text{in}}} - w^\star {w^\star}^\top)$ to capture structure in the invariant feature that the variance is less along the latent predictive direction $w^\star$. Note that the spurious feature is completely predictive of the label in the source data, and is distributed as spherical Gaussian in the target data with $\Sigma_{\text{sp}} = \sigma_{\text{sp}}^2 \mathbf{I}_{d_{\text{sp}}}$.

Our distribution shift setting bears similarities but also exhibits important differences (discussed below). For mathematical convenience, we assume access to infinite unlabeled data and hence replace the empirical quantities over unlabeled data with their population counterpart. For SSL, we sample finite labeled and infinite unlabeled data from $\mathrm{P_T}$ where spurious features are absent (to exclude easy-to-generalize features). For UDA, we further assume access to infinite labeled data from the source. Note that due to distribution shift, population access of labeled data doesn't trivialize the problem as "ERM" on infinite labeled source data does not necessarily achieve optimal performance on the target.

Methods and objectives Recall from Section 1.1 that we learn linear classifiers h over features extractors Φ. We consider linear feature extractor i.e. Φ is a matrix in $\mathbb{R}^{d \times k}$ and the linear layer $h : \mathbb{R}^k \to \mathbb{R}$ with a prediction as $\mathrm{sgn}(h^\top \Phi x)$. We use the exponential loss $\ell(f(x), y) = \exp(-y f(x))$.

Self-training. ST performs ERM in the first stage using labeled data from the source, and then subsequently updates the head h by iteratively generating pseudolabels on the unlabeled target:

$$\mathcal{L}_{\text{st}}(h; \Phi) := \mathbb{E}_{\mathrm{P_T}(x)} \ell(h^\top \Phi x, \mathrm{sgn}(h^\top \Phi(x))) \qquad \text{Update: } h^{t+1} = \frac{h^t - \eta \nabla_h \mathcal{L}_{\text{st}}(h^t; \Phi)}{\left\| h^t - \eta \nabla_h \mathcal{L}_{\text{st}}(h^t; \Phi) \right\|_2} \tag{5}$$

For ERM and ST, we train both h and Φ (equivalent to Φ being identity and training a linear head).

Contrastive pretraining. We obtain $\Phi_{\text{cl}} := \arg\min_\Phi \mathcal{L}_{\text{cl}}(\Phi)$ by minimizing the Barlow Twins objective (Zbontar et al., 2021), which prior works have shown is also equivalent to spectral contrastive and non-contrastive objectives (Garrido et al., 2022; Cabannes et al., 2023). Given probability distribution $\mathrm{P_A}(a \mid x)$ for input x, and marginal $\mathrm{P_A}$, we consider a constrained form of Barlow Twins in (6) which enforces features of "positive pairs" a_1, a_2 to be close while ensuring feature diversity. We assume a strict regularization ($\rho = 0$) for the theory arguments in the rest of the paper, and in App. G.2 we prove that all our claims hold for small ρ as well. For augmentations, we scale the magnitude of each co-ordinate uniformly by an independent amount, i.e., $a \sim \mathrm{P_A}(\cdot \mid x) = \mathbf{c} \odot x$, where $\mathbf{c} \sim \mathrm{Unif}[0, 1]^d$. We try to mirror practical settings where the augmentations are fairly "generic", not encoding information about which features are invariant or spurious, and hence perturb all features symmetrically.

$$\mathcal{L}_{\text{cl}}(\Phi) := \mathbb{E}_{x \sim \mathrm{P_U}} \mathbb{E}_{a_1, a_2 \sim \mathrm{P_A}(\cdot \mid x)} \left\| \Phi(a_1) - \Phi(a_2) \right\|_2^2 \ \text{ s.t. } \ \left\| \mathbb{E}_{a \sim \mathrm{P_A}} \left[\Phi(a) \Phi(a)^\top \right] - \mathbf{I}_k \right\|_F^2 \leqslant \rho \tag{6}$$

Keeping the Φ_{cl} fixed, we then learn a linear classifier h_{cl} over Φ_{cl} to minimize the exponential loss on labeled source data (refer to as *linear probing*). For STOC, keeping the Φ_{cl} fixed and initializing the linear head with the CL linear probe (instead of source only ERM), we perform ST with (5).

Example G.1. For the setup in (4), we choose $\gamma = 0.5$, $\sigma_{\text{sp}}^2 = 1.$, and $\sigma_{\text{in}}^2 = 0.05$ with $d_{\text{in}} = 5$ and $d_{\text{sp}} = 20$ for our running example. $\gamma/\sqrt{d_{\text{sp}}}$ controls signal to noise ratio in the source such that spurious feature is easy-to-learn and the invariant feature is harder-to-learn. σ_2 controls the noise in target which we show later is critical in unlearning the spurious feature with CL.

Why is our simplified setup interesting? In our setup, x_{in} is the hard to learn feature that generalizes from source to target. The hardness of learning this feature is determined by the value of the margin γ and how it compares with size of the spurious feature ($\sqrt{d_{\text{sp}}}$). Since, $\gamma/\sqrt{d_{\text{sp}}}$ is small in our setup, x_{in} is much harder to learn on source data (even with population access) compared to the spurious feature x_{sp} which generalizes poorly from source to target. These two types of features have been captured in similar analysis on spurious correlations (Sagawa et al., 2020; Nagarajan et al., 2020) since it imitates pitfalls emanating from the presence of spurious features in real world datasets (*e.g.*, the easy to learn background feature in image classification problems). While this setup is simple, it is also expressive enough to elucidate both self-training and contrastive learning behaviors we observe in real world settings. Specifically, it captures the separation results we observe in Sec. 2.

Differences of our setup with prior works. While our distribution shift settings bears the above similarities it also has important differences with works analyzing self-training and contrastive pretraining individually. Chen et al. (2020b) analyze the iterative nature of self-training algorithm, where the premise is that we are given a classifier that not only has good performance on source data but in addition does not rely too much on the spurious feature. Under the strong condition of small norms along the spurious feature, they show that self-training can provably unlearn this small dependence when the target data along the spurious feature is random noise. This assumption is clearly violated in setups where the spurious correlation is strong (as in our toy setup), *i.e.*, the dependence on the spurious feature is rather large (much larger than that on the invariant feature) for any classifier that is trained directly on source data. Consequently, we show the need for "good" pretrained representations from contrastive pretraining over which if we train a linear predictor (using source labeled data), it will provably have a reduced "effective" dependence on the spurious feature.

Using an augmentation distribution similar to ours, Saunshi et al. (2022) carried out contrastive pretraining analysis with the backbone belonging to a capacity constrained function class (similar analysis also in (HaoChen et al., 2022)). Our setup differs from this in two key ways: (i) we specifically consider a distribution shift from source to target. Unlike their setting, it is not sufficient to make augmentations consistent with ground truth labels, since the predictor that uses just the spurious feature also assigns labels consistent with both ground truth predictions and augmentations on the source data; and (ii) our augmentation distribution assumes no knowledge of the invariant feature, which is why we augment all dimensions uniformly, as opposed to selectively augmenting a set of dimensions. In other words, we assume no knowledge of the structure of the optimal target predictor. For *e.g.*, if we had knowledge of the spurious dimensions we could have just selectively augmented those. Assuming knowledge of these perfect augmentations is not ideal for two reasons: (a) it makes the problem so easy that just training an ERM model on source data with these augmentations would already yield a good target predictor (which rarely happens in practice); and (b) in real-world datasets perfect augmentations for the downstream task are not known. Hence, we stick to generic augmentations in our setup.

G.2. Discussion on self-training and contrastive learning objectives

Method	**UDA Setup**	**SSL Setup**
ERM:	$h_{\mathrm{erm}} = \arg\min_h \mathbb{E}_{\mathrm{P_S}} \ell(h(x), y)$	$h_{\mathrm{erm}} = \arg\min_h \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$ $\{(x_i, y_i)\}_{i=1}^{n} \sim \mathrm{P_T}^n$
ST:	Starting from h_{erm} optimize over h (to get h_{st}): $\mathbb{E}_{\mathrm{P_T}(x)} \ell(h(x), \mathrm{sgn}(h(x)))$	Starting from h_{erm} optimize over h (to get h_{st}): $\mathbb{E}_{\mathrm{P_T}(x)} \ell(h(x), \mathrm{sgn}(h(x)))$
CL:	$\Phi_{\mathrm{cl}} = \arg\min_\phi \mathcal{L}_{\mathrm{cl}}(\Phi)$ Use $(\mathrm{P_S}(x) + \mathrm{P_T}(x))/2$ for $\mathcal{L}_{\mathrm{cl}}(\Phi)$ $h_{\mathrm{cl}} = \arg\min_h \mathbb{E}_{\mathrm{P_S}} \ell(h \circ \Phi_{\mathrm{cl}}(x), y)$	$\Phi_{\mathrm{cl}} = \arg\min_\phi \mathcal{L}_{\mathrm{cl}}(\Phi)$ Use $\mathrm{P_T}(x)$ for $\mathcal{L}_{\mathrm{cl}}(\Phi)$ $h_{\mathrm{cl}} = \arg\min_h \frac{1}{n}\sum_{i=1}^{n} \ell(h \circ \Phi_{\mathrm{cl}}(x_i), y_i)$
STOC:	Starting from h_{cl} optimize over h (to get h_{stoc}): $\mathbb{E}_{\mathrm{P_T}(x)} \ell(h \circ \Phi_{\mathrm{cl}}(x), \mathrm{sgn}(h \circ \Phi_{\mathrm{cl}}(x)))$	Starting from h_{cl} optimize over h (to get h_{stoc}): $\mathbb{E}_{\mathrm{P_T}(x)} \ell(h \circ \Phi_{\mathrm{cl}}(x), \mathrm{sgn}(h \circ \Phi_{\mathrm{cl}}(x)))$

Table 8: **Description of methods for SSL vs. UDA**: For each method we provide exact objectives used for experiments and analysis in the SSL and UDA setups (pertaining to Sec. 3).

In text we will describe our objectives and methods for the UDA setup. In Table 8 we constrast the differences in the methods and objectives for SSL and UDA setups. Recall from Section 1.1 that we learn linear classifiers h over features extractors Φ. We consider linear feature extractor i.e. Φ is a matrix in $\mathbb{R}^{k \times d}$. For mathematical convenience, we assume access to infinite unlabeled data and hence replace the empirical quantities over unlabeled data with their population counterpart. In the UDA setting, we further assume access to infinite labeled data from the source. Note that due to distribution shift between source and target, "ERM" on infinite labeled data from the source does not necessarily achieve optimal performance on the target. For binary classification, we assume that the linear layer h maps features to a scalar in $\mathbb{R}$ such that the prediction is $\mathrm{sgn}(h^\top \Phi x)$. We use the exponential loss $\ell(f(x), y) = \exp(-y f(x))$ as the classification loss.

Contrastive pretraining. We obtain $\Phi_{\mathrm{cl}} := \arg\min_\Phi \mathcal{L}_{\mathrm{cl}}(\Phi)$ by minimizing the Barlow Twins objective (Zbontar et al., 2021), which prior works have shown is also equivalent to spectral contrastive and non-contrastive objectives (Garrido et al., 2022; Cabannes et al., 2023). In Sec. 3, we consider a constrained form of Barlow Twins in (6) which enforces representations of different augmentations a_1, a_2 of the same input x to be close in representation space, while ensuring feature diversity by staying in the constraint set. We assume a strict constraint on regularization ($\rho = 0$) for the theoretical arguments in the rest of the main paper. In App. H.6.2 we prove that all our claims hold for small ρ as well. In (7), we redefine the pretraining objective with a regularization term (instead of a constraint set) where κ controls the strength of the regularization term, with higher values of κ corresponding to stronger constraints on feature diversity. We then learn a linear classifier h_{cl} over Φ_{cl} to minimize the exponential loss on labeled source data.

$$\mathcal{L}_{\mathrm{cl}}(\Phi) \;:=\; \mathbb{E}_{x \sim \mathrm{P_U}} \mathbb{E}_{a_1, a_2 \sim \mathrm{P_A}(\cdot|x)} \, \|\Phi(a_1) - \Phi(a_2)\|_2^2 \;+\; \kappa \cdot \left\| \mathbb{E}_{a \sim \mathrm{P_A}} \left[\Phi(a)\Phi(a)^\top\right] - \mathbf{I}_k \right\|_F^2 \tag{7}$$

Augmentations. Data augmentations play a key role in contrastive pre-training (and also as we see later, state-of-the-art self-training variants like FixMatch). Given input $x \in \mathcal{X}$, let $\mathrm{P_A}(a \mid x)$ denote the distribution over its augmentations, and $\mathrm{P_A}$ denote the marginal distribution over all possible augmentations. We use the following simple augmentations where we scale the magnitude of each co-ordinate by a uniformly independent amount, *i.e.*,

$$a \sim \mathrm{P_A}(\cdot \mid x) \equiv c \odot x \quad \text{where,} \quad c \sim \mathrm{Unif}[0,1]^d. \tag{8}$$

The performance of different methods heavily depends on the assumptions we make on augmentations. We try to mirror practical settings where the augmentations are fairly "generic", not encoding any information about which features are invariant or spurious, and hence perturb all features symmetrically.

Self-training. ST performs ERM in the first stage using labeled data from the source, and then subsequently updates the head h by iteratively generating pseudolabels on the unlabeled target:

$$\mathcal{L}_{\mathrm{st}}(h; \Phi) \;:= \mathbb{E}_{\mathrm{P_T}(x)} \ell(h^\top \Phi x, \mathrm{sgn}(h^\top \Phi(x))) \qquad \text{Update:} \;\; h^{t+1} = \frac{h^t - \eta \nabla_h \mathcal{L}_{\mathrm{st}}(h^t; \Phi)}{\|h^t - \eta \nabla_h \mathcal{L}_{\mathrm{st}}(h^t; \Phi)\|_2} \tag{9}$$

For convenience, we keep the feature backbone Φ fixed across the self-training iterations and only update the linear head on the pseudolabels.

STOC(Self-training after contrastive learning). Finally, we can combine the two unsupervised objectives where we do the self-training updates(5) with $h_0 = h_{\mathrm{cl}}$ and $\Phi_0 = \Phi_{\mathrm{cl}}$ starting with the contrastive learning model rather than just source-only ERM. Here, we only update h and fix Φ_{cl}.

G.3. Additional empirical results in our simplified setup

We conduct two ablations on the hyperparameters for contrastive pretraining. First, we vary the dimensionality k of the linear feature extractor $\Phi \in \mathbb{R}^{k \times d}$. Second, we vary the regularization strength κ that enforces feature diversity in the Barlow Twins objective (7). In Figure 5 we plot these ablations in the UDA setup.

Varying feature dimension. We find that CL recovers the full set of predictive features (*i.e.* both spurious and invariant) only when k is large enough (Figure 5*(left)*). Since the dimensionality of the true feature is 5 in our Example 1, reducing k below the true feature dimension hurts CL. Once k crosses a certain threshold, CL features completely capture the projection of the invariant feature w_{in}. After this point, it amplifies the component along w_{in}. It retains the amplification over the spurious feature w_{sp} even as we increase k. This is confirmed by our finding that further increasing k does not hurt CL performance. This is also inline with our theoretical observations, where we find that for suitable $w^\star$, the subspace spanned by w_{in} and w_{sp} are contained in a low rank space (as low as rank 2) of the contrastive representations (Theorem H.3). Once CL has amplified the dependence along w_{in} STOC improves over CL by unlearning any remaining dependence on the spurious w_{sp}. The above arguments for the CL trend also explain why the performance of STOC continues to remain $\approx 100\%$ as we vary k.

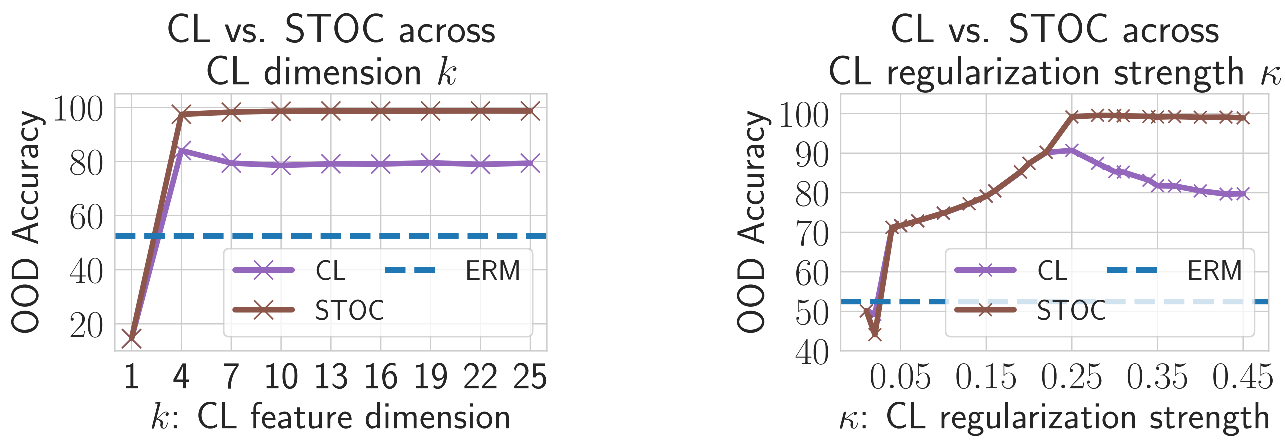


Figure 5: **Ablations on pretraining hyperparameters:** In the UDA setup we plot the performance of CL and STOC as we vary two pretraining hyper-parameters: *(left)* the output dimension (k) of the feature extractor Φ; and *(right)* the strength (κ) of the regularizer in the Barlow Twins objective in (7). While ablating on k we fix $\kappa = 0.5$, and while ablating on κ we fix $k = 10$. Other problem parameters are taken from Example 1.

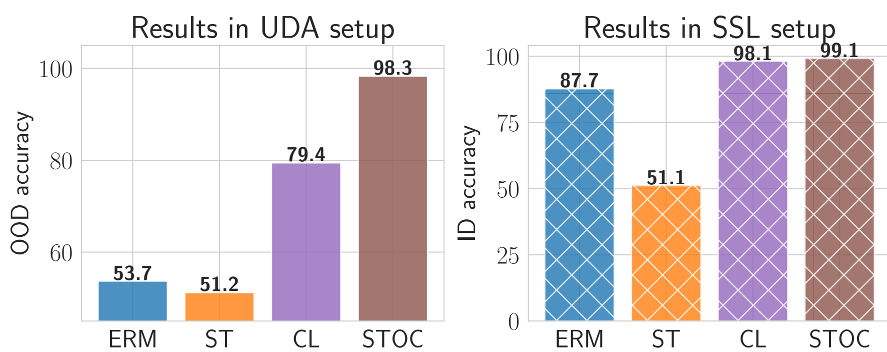


Figure 6: **Results with linear backbone:** We plot the OOD accuracy for ERM, CL, ST and STOC in the UDA setup and ID accuracy in the SSL setup when the feature extractor Φ is a linear network. Note, that the feature extractor is still fixed during CL and STOC.

Varying regularization strength. In our main theoretical arguments we consider the constrained form of the Barlow Twins objective (6) with a strict constraint of $\rho = 0$ (we relax this theoretically as well, see H.6.2). For our experiments, we optimize the regularized version of this objective (7), where the constraint term now appears as a regularizer which enforces feature diversity, *i.e.* the features learned through contrastive pretraining span orthogonal parts of the input space (as governed under the metric defined by augmentation covariance matrix Σ_A). If κ is very low, then trivial solutions exist for the Barlow Twins objective. For *e.g.*, $\phi \approx \mathbf{0}$ (zero vector) achieves very low invariance loss. When $\kappa < 0.05$, we find that CL recovers these trivial solutions (Figure 5*(right)*). Hence, both CL and STOC perform poorly. As we increase κ the performance of both CL and STOC improve, mainly because the features returned by Φ_{cl} now comprise of the predictive directions w_{in} and w_{sp}, as predictive by our theoretical arguments for $\rho = 0$ (which corresponds to large κ). On the other hand, when κ is too high optimization becomes hard since κ directly effects the Lipschitz constant of the loss function. Hence, the performance of CL drops by some value. Note that this does not effect the performance of STOC since CL continues to amplify w_{in} over w_{sp} even if it is returning suboptimal solutions with respect to the optimization loss of the pretraining objective.

G.4. Reconciling Practice: Experiments with deep networks in toy setup

In this section we delve into the details of Sec. H.4, *i.e.*, we analyze performance of different methods when we make some design choices that imitate practice. First, we look at experiments involving a deep non-linear backbone Φ. Here, the non-linear Φ is learned during contrastive pretraining and fixed for CL and STOC. Then, we investigate trends when we continue to propagate gradients onto Φ during STOC (we call this full-finetuning). Unlike previous cases, this allows features to be updated.

Results with non-linear feature extractor Φ. In Fig. 7 we plot the performance of the four methods when we use a non-linear feature extractor during contrastive pretraining. This feature extractor is a one-hidden layer neural network (hidden dimension is 500) with ReLU activations. We find that the trends observed with linear backbones in Fig. 6 are also replicated with the non-linear one. Specifically, we note that STOC improves over CL under distribution shifts, whereas CL is already close to optimal when there are no distribution shifts. We also see that CL and ST individually are subpar. In SSL, we see a huge drop in the performance of ST (over ERM) mainly because we only fit on pseudolabels during ST. This is different from practice where we continue to optimize loss on labeled data points while fitting the pseudolabels. Consequently, when we continue to optimize performance on source labeled data the performance of ST in SSL setup is improves from $51.1\% \rightarrow 72.6\%$.

Results with full fine-tuning. Up till this point, we have only considered the case (for both SSL and UDA) where we fix the contrastive learned features when running CL and STOC, *i.e.*, we only optimized the linear head h. Now, we shall consider the setting where gradients are propagated to Φ during STOC. Note that we still fix the representations for training the linear head during CL. Results for this setting are in Figure 8. We show two interesting trends that imitate real world behaviors.

STOC benefits from augmentations during full-finetuning: In the UDA setup we find that ST while updating Φ_{cl} can hurt due to overfitting issues when training with the finite sample of labeled and unlabeled data (drop by $> 7\%$ over CL). This is due to overfitting on confident but incorrect pseudolabels on target data. This can exacerbate components along spurious feature w_{sp} from source. One reasoning behind this is that deep neural networks can perfectly memorize them on finite unlabeled target data (Zhang et al., 2017). Heuristics typically used in practice (*e.g.* in FixMatch (Sohn et al., 2020)) help avoid overfitting on incorrect pseudolabels: (i) confidence thresholding; to pick confident pseudolabel examples; (ii) pseudolabel a different augmented input than the one on which the self-training loss is optimized; and (iii) optimize source loss with labeled data simultaneously when fitting pseudolabels. Intuitively, thresholding introduces a curriculum where we only learn confident examples in the beginning whose pseudolabels are mainly determined by component along the invariant feature w_{in}. Augmentations prevent the neural network from memorizing incorrect pseudolabels and optimizing source loss prevents forgetting of features learned during CL. When we implement these during full-finetuning in STOC we see that STOC now improves over CL (by $> 20\%$).

Can we improve contrastive pretraining features during STOC? We find that self-training can also improve features learned during contrastive pretraining when we update the full backbone during STOC (see Figure 8(*right*)). Specifically, in the SSL setup we find that STOC can now improve substantially over CL. Recall, that when we fixed Φ_{cl} this was not possible (see

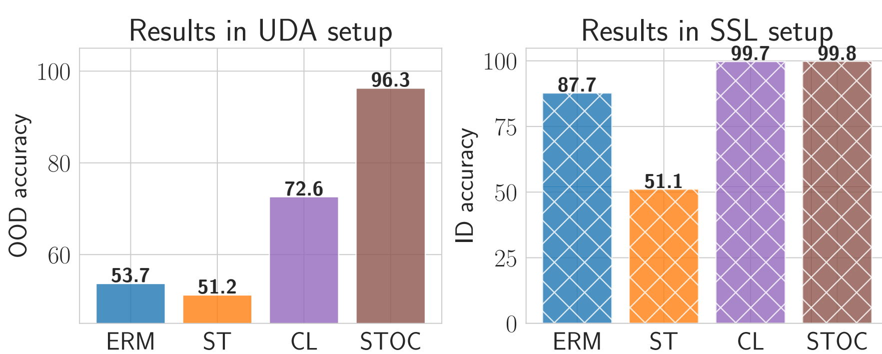


Figure 7: **Results with non-linear backbone:** We plot the OOD accuracy for ERM, CL, ST and STOC in the UDA setup and ID accuracy in the SSL setup when the feature extractor Φ is a non-linear one-hidden layer network with ReLU activations. Note, that the feature extractor is still fixed during CL and STOC.

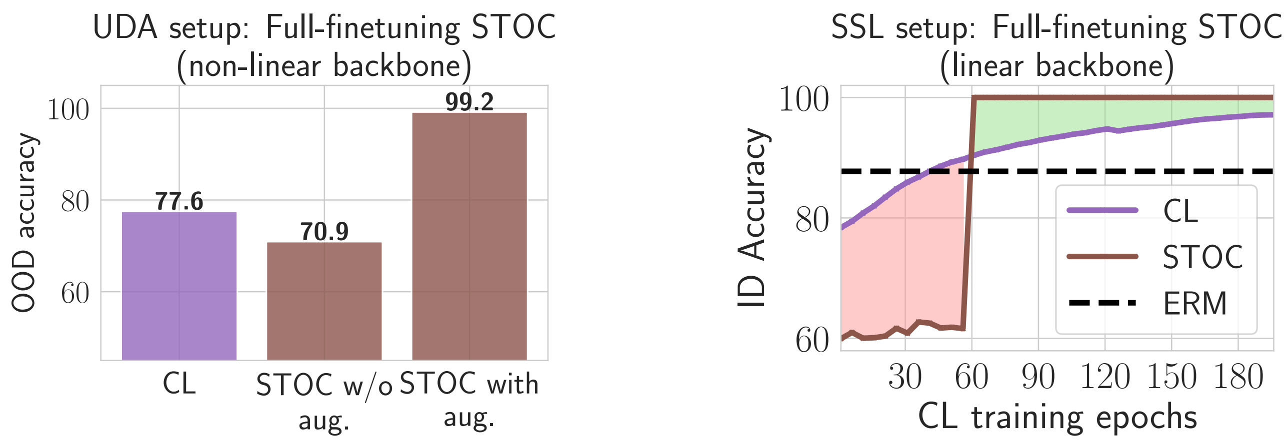


Figure 8: **Finetuning the contrastive representations during STOC:** We propagate gradients to the feature backbone Φ when running STOC algorithm. Note that CL still fixes the contrastive representations when learning a fixed linear head over it. On the *(left)* we show results in UDA setup where we compare the performance of STOC with and without augmentations (along with other practical design choices like confidence thresholds and continuing to optimize source loss as done in FixMatch) when the feature backbone is non-linear. On the *(right)* we show results for STOC and CL in the SSL setup when the feature backbone is linear.

H.8 and Fig. 2(b)). This is mainly because STOC can now improve performance beyond just recovering the generalization gap for the linear head (which is typically small). This feature improvement is observed even when we fully finetune a linear feature extractor. Similar trends are also observed with the non-linear backbone. But, it becomes harder to identify a good stopping criterion for CL training. Thus, it remains unclear if STOC and CL have complementary benefits for feature learning in UDA or SSL settings. Investigating this is an interesting avenue for future work.

H. Theoretical Results from Sec. 3

H.1. Conditions for Success and Failure of Self-training over ERM from Scratch

In our results on Example G.1, we observe that performing ST after ERM yields a classifier with near-random target accuracy. In Theorem H.1, we characterize conditions under which ST fails and succeeds.

Theorem H.1 (Informal; Conditions for success and failure of ST over ERM)**.** *The target accuracy of ERM classifier, is given by* $0.5 \cdot \text{erfc}\left(-\gamma^2/(\sqrt{2d_{\text{sp}}}\cdot\sigma_{\text{sp}})\right)$*. Then for* $\sigma_{\text{sp}} > \sigma_0$*, ST performed in the second stage yields: (i) a classifier with* ≈ 0.5 *target accuracy when* $\gamma/d_{\text{sp}} < c_1 \cdot \sigma_{\text{sp}}$*; and (ii) a classifier with near-perfect target accuracy when* $\gamma/d_{\text{sp}} \gg c_1 \cdot \sigma_{\text{sp}}$ *for some constant* c_1*.*

The informal theorem above abstracts the exact dependency of $\gamma, \sigma_{\text{sp}}$, and d_{sp} for the success and failure of ST over ERM. Our analysis highlights that while ERM learns a perfect predictor along $w_{\text{in}}{=}[w^\star, 0, ..., 0]^\top$ (with norm γ), it also learns to depend on $w_{\text{sp}} = [0, ..., 0, \mathbf{1}_{d_{\text{sp}}}/\sqrt{d_{\text{sp}}}]^\top$ with norm $\sqrt{d_{\text{sp}}}$ because of the perfect correlation of x_{sp} with labels on the source. Our conditions depict that when the γ/d_{sp} is sufficiently smaller than σ_{sp}, then ST continues to erroneously enhance its reliance on the x_{sp} feature for target prediction, resulting in near-random target performance. Conversely, when γ/d_{sp} is much larger than σ_{sp}, the signal in x_{in} is correctly used for predictor on the majority of target points, and ST eliminates the x_{sp} dependency, converging to an optimal target classifier.

Our proof analysis shows that if the ratio of the norm of the classifier along in the direction of $w^\star$ is smaller than w_{sp} by a certain ratio then the generated pseudolabels (incorrectly) use x_{sp} for its prediction further increasing the component along w_{sp}. Moreover, normalization further diminishes the reliance along $w^\star$, culminating in a near-random performance. The opposite occurs when the ERM classifier achieves a signal along $w^\star$ that is sufficiently stronger than along w_{sp}. Upon substituting the parameters used in Example G.1, the ERM and ST performances as determined by Theorem H.1 align with our empirical results, notably, ST performance on target being near-random.

H.2. CL Captures Both Features But Amplifies Invariant Over Spurious Features

Recall, minimizing the contrastive loss in (6) gives us Φ_{cl}, for which we derive a closed form expression in Proposition H.2 that holds generally for any linear backbone and augmentation distribution.

Proposition H.2 (Barlow Twins solution)**.** *The solution for* (6) *is* $U_k^\top \Sigma_{\mathsf{A}}^{-1/2}$ *where* U_k *are the top* k *eigenvectors of* $\Sigma_{\mathsf{A}}^{-1/2}\,\widetilde{\Sigma}\,\Sigma_{\mathsf{A}}^{-1/2}$. *Here,* $\Sigma_{\mathsf{A}} := \mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[aa^\top]$ *is the covariance over augmentations, and* $\widetilde{\Sigma} := \mathbb{E}_{x\sim \mathrm{P}_{\mathsf{U}}}[\widetilde{a}(x)\widetilde{a}(x)^\top]$ *is the covariance matrix of mean augmentations* $\widetilde{a}(x) := \mathbb{E}_{\mathrm{P}_{\mathsf{A}}(a|x)}[a]$.

Intuitively, the above result captures the effect of augmentations through the matrix U_k. If there were no augmentations, then $\Sigma_{\mathsf{A}} = \widetilde{\Sigma}$, implying that U_k could then be any random orthonormal matrix. On the other hand if augmentation distributions change prevalent covariances in the data, *i.e.*, Σ_{A} is very different from original feature covariance on actual data, the matrix U_k would bias the CL solution towards directions that capture significant variance in mean augmentations but only if augmentations do not scale the variance along it by a lot—precisely the directions with low invariance loss. In the final solution $U_k^\top \Sigma_{\mathsf{A}}^{-1/2}$, while the invariance loss in (6) determines U_k, the constraint in (6) determines the norm along each direction which is corrected once U_k is scaled by $\Sigma_{\mathsf{A}}^{-1/2}$.

Based on this we can conjecture, that CL would learn components along both invariant w_{in} and spurious w_{sp} components because: (i) these directions explain a large fraction of variance in the raw data; (ii) augmentations that randomly scale down dimensions would not add a lot of variance along w_{sp} and w_{in} as compared to noise directions in their null space. But, since the spurious feature is random on target, the variance along w_{sp} in target would be much higher under augmentations as compared to that along the invariant w_{in}. Thus, when CL is done on the union of source and target unlabeled data, it would amplify w_{in} over w_{sp}. For $w^\star = \mathbf{1}_{d_{\text{in}}}/\sqrt{d_{\text{in}}}$, we formalize this intuition in Theorem H.3. While we do this for mathematical convenience in trying to analyze claims tightly, our results in Sec. 3.1 hold for the general case of any $w^\star$ (for discussion on this, see App. G.1).

Theorem H.3 (Informal; CL recovers both invariant w_{in} and spurious w_{sp} but amplifies w_{in})**.** *For* $w^\star = \mathbf{1}_{d_{\text{in}}}/\sqrt{d_{\text{in}}}$, *the CL solution* $\Phi_{\text{cl}}=[\phi_1, \phi_2, ..., \phi_k]$ *satisfies* $\phi_j^\top w_{\text{in}} = \phi_j^\top w_{\text{sp}} = 0\ \forall j \geqslant 3$, $\phi_1 = c_1 w_{\text{in}} + c_3 w_{\text{sp}}$ *and* $\phi_2 = c_2 w_{\text{in}} + c_4 w_{\text{sp}}$. *For bounded* $\gamma/\sqrt{d_{\text{sp}}}$ *and* σ_{sp}, *the signal along* w_{in} *is amplified,* i.e., *for some small* $\epsilon > 0$, $|c_2/c_4| \geqslant (1-\epsilon)\sqrt{d_{\text{sp}}}/\gamma$ *and* $5\gamma/\sqrt{d_{\text{sp}}} \leqslant c_1/c_3 \leqslant 20\gamma/\sqrt{d_{\text{sp}}}$.

Based on our intuition above, Theorem H.3 first conveys that CL recovers components along both w_{in} and w_{sp} through ϕ_1, ϕ_2 where it increases the norm along w_{in} more than w_{sp}. We can see this because the margin separating labeled points along w_{in} is now amplified by a factor of $|c_2/c_4| = \Omega(\sqrt{d_{\text{sp}}}/\gamma)$ in ϕ_1 and $c_1/c_3 \geqslant 2\gamma$ in ϕ_2, as compared to the same margin on source distribution. Naturally, this will improve the target performance of a linear predictor trained over CL representations. At the same time, we also see that in ϕ_1, the component along w_{sp} is still significant (c_1/c_3 is upper bounded). This is because, while the random noise along w_{sp} in target is amplified by augmentations, the variance induced by augmentations along w_{sp} in source is still very small. Due the remaining components along w_{sp}, the target performance for CL can remain less than ideal. Both the above arguments on target performance are captured in Corollary H.4.

Corollary H.4 (Informal; CL improves OOD error over ERM but is still imperfect)**.** *Under the conditions of Theorem H.3 the target accuracy of CL is at least* $0.5 \cdot \text{erfc}\left(-c_1/(\omega c_3) \cdot \gamma/\left(\sqrt{2d_{\text{sp}}}\cdot\sigma_{\text{sp}}\right)\right)$, *and at most* $\leqslant 0.5 \cdot \text{erfc}\left(-c_1/c_3 \cdot \gamma/\left(\sqrt{2d_{\text{sp}}}\cdot\sigma_{\text{sp}}\right)\right)$. *Note that since* $c_1/c_3 \geqslant 5\gamma/\sqrt{d_{\text{sp}}}$, *the lower bound is strictly better than ERM when* $1 \leqslant \omega \leqslant 5$.

While Φ_{cl} is still not ideal for linear probing, in the next part we will see how Φ_{cl} can instead be sufficient for subsequent self-training to unlearn the remaining components along spurious features.

H.3. Improvements with Self-training Over Contrastive Learning

The result in the previous section highlights that while CL may improve over ERM, the linear probe continues to depend on the spurious feature. Next, we characterize the behavior STOC. Recall, in the ST stage, we iteratively update the linear head with (5) starting with the CL backbone and head.

Theorem H.5 (Informal; ST improves over CL)**.** *Under the conditions of Theorem H.3, the target accuracy of ST over CL is lower bounded by* $0.5 \cdot \text{erfc}\left(-|c_2/c_4| \cdot \gamma/(\sqrt{2}\sigma_2)\right) \approx 0.5 \cdot \text{erfc}\left(-\sqrt{d_2}/(\sqrt{2}\sigma_2)\right)$ *where* c_2 *and* c_4 *are the coefficients of feature* ϕ_2 *along* $w^\star$ *and* w_{sp} *learned by BT.*

The above theorem states that when $\sqrt{d_2}/\sigma_2 \ll 1$ the target accuracy of ST over CL is close to 1. In Example G.1, the lower bound of the accuracy of ST over CL is $\text{erfc}\left(-\sqrt{10}\right) \approx 2$ showing near-perfect target generalization. Recall that Theorem H.4 shows that CL yields a linear head that mainly depends on both the invariant direction w^* and the spurious direction w_{sp}. At initialization, the linear head trained on the CL backbone has negligible dependence on ϕ_2 (under conditions in Theorem H.4). Building on that, the analysis in Theorem H.5 captures that ST gradually reduces the dependence on w_{sp} by learning a linear head that has a larger reliance on ϕ_2, which has a higher "effective" margin on the target, thus increasing overall dependency on w^*.

Theoretical comparison with SSL. Our analysis until now shows that linear probing with source labeled data during CL picks up features that are more predictive of source label under distribution shift, leaving a significant room for improvement on OOD data when self-trained further. In UDA, the primary benefit of ST lies in picking up the features with a high "effective" margin on target data that are not picked up by linear head trained during CL. In contrast, in the SSL setting, the limited ID labeled data may provide enough signal in picking up high-margin features which are predictive on ID data, leaving little to no room for improvement for further ST. We formalize this intuition in App. H.

H.4. Reconciling Practice: Implications for Deep Non-Linear Networks

In this section, we experiment with deep non-linear backbone (*i.e.*, Φ_{cl}). When we continue to fix Φ_{cl} during CL and STOC, the trends we observed with linear networks in Sec. 3.1 continue to hold. We then perform full fine-tuning with CL and STOC, i.e., propagate gradients even to Φ_{cl}, as commonly done in practice. We present key takeaways here but detailed experiments are in App. G.4.

Benefits of augmentation for self-training. ST while updating Φ_{cl} can hurt due to overfitting issues when training with the finite sample of labeled and unlabeled data (drop by >10% over CL). This is due to the ability of deep networks to overfit on confident but incorrect pseudolabels on target data (Zhang et al., 2017). This exacerbates components along w_{sp} and we find that augmentations (and other heuristics) typically used in practice (*e.g.* in FixMatch (Sohn et al., 2020)) help avoid overfitting on incorrect pseudolabels.

Can ERM and ST over contrastive pretraining improve features? We find that self-training can also slightly improve features when we update the backbone with the second stage of STOC and when the CL backbone is early stopped sub-optimally (*i.e.* at an earlier checkpoint in Fig. 2(b)). This feature finetuning can now widen the gap between STOC and CL in SSL settings, as compared to the linear probing gap (as in 2). This is because STOC can now improve performance beyond just recovering the generalization gap for the linear head (which is typically small). However, STOC benefits are negligible when CL is not early stopped sub-optimally, *i.e.*, trained till convergence. Thus, it remains unclear if STOC and CL have complementary benefits for feature learning in UDA or SSL settings. Investigating this is an interesting avenue for future work.

Jumping into formal proofs and analysis. The above discussion concludes our empirical findings in our simplified setup. Next, we jump into the proofs for the theorems introduced in previous subsections. Before that, recall from Section 1.1 that we learn linear classifiers h over features extractors Φ. We consider linear feature extractor i.e. Φ is a matrix in $\mathbb{R}^{d\times k}$ and the linear layer $h : \mathbb{R}^k \to \mathbb{R}$ with a prediction as $\text{sgn}(h^\top \Phi x)$. We use the exponential loss $\ell(f(x), y) = \exp\left(-yf(x)\right)$.

H.5. Analysis of ERM and ST: Formal Statement of Theorem H.1

For ERM and ST, we train both h and Φ. This is equivalent to $\Phi = I_{d\times d}$ being identity and training a linear head h. Recall that the ERM classifier is obtained by minimizing the population loss on labeled source data:

$$h_{\text{ERM}} = \arg\min_h \mathbb{E}_{(x,y)\sim \mathrm{P_S}}\left[\ell(x,y)\right] . \tag{10}$$

We split Theorem H.1 into Theorem H.6 and Theorem H.7. Before we characterize the ERM solution, we recall some additional notation. Define $w_{\text{in}}{=}[w^\star, 0, ..., 0]^\top$, and $w_{\text{sp}} = [0, ..., 0, \mathbf{1}_{d_{\text{sp}}}/\sqrt{d_{\text{sp}}}]^\top$. The following proposition characterizes h_{ERM} and 0-1 error of the classifier on target:

Theorem H.6 (ERM classifier and its error on target)**.** *ERM classifier obtained as in* (10) *is given by*

$$\frac{h_{\mathit{ERM}}}{\|h_{\mathit{ERM}}\|_2} = \frac{\gamma \cdot w_{\text{in}} + \sqrt{d_{\text{sp}}} \cdot w_{\text{sp}}}{\sqrt{\gamma^2 + d_{\text{sp}}}} .$$

The target accuracy of h_{ERM} *is given by* $0.5 \cdot \text{erfc}\left(-\gamma^2/(\sqrt{2d_{\text{sp}}}\cdot\sigma_{\text{sp}})\right)$.

Proof. To prove this theorem, we first derive a closed-form expression for the ERM classifier and then use Lemma J.9 to derive its 0-1 error on target. For Gaussian data with the same covariance matrices for class conditional $\mathrm{P_S}(x|y=1)$ and $\mathrm{P_S}(x|y=0)$, Bayes decision rule is given by the Fisher's linear discriminant direction (Chapter 4; Bishop (2006)):

$$h(x) = \begin{cases} 1, & \text{if } h^\top x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $h = 2 \cdot \gamma(w_{\text{in}}) + 2 \cdot \sqrt{d_{\text{sp}}}(w_{\text{sp}})$. Plugging h in Lemma J.9 we get the desired result. □

ST performs ERM in the first stage using labeled data from the source, and then subsequently updates the head h by iteratively generating pseudolabels on the unlabeled target:

$$\mathcal{L}_{\text{st}}(h) \;:= \mathbb{E}_{\mathrm{P_T}(x)} \ell(h^\top x, \operatorname{sgn}(h^\top x))\,. \tag{11}$$

Starting with $h^0_{\text{ST}} = {h_{\text{ERM}}}/{\|h_{\text{ERM}}\|_2}$ (the classifier obtained with ERM) we perform the following iterative procedure for self-training:

$$h^{t+1}_{\text{ST}} = \frac{h^t_{\text{ST}} - \eta \nabla_h \mathcal{L}_{\text{st}}(h^t_{\text{ST}})}{\|h^t_{\text{ST}} - \eta \nabla_h \mathcal{L}_{\text{st}}(h^t_{\text{ST}})\|_2} \tag{12}$$

Next, we characterize ST solution:

Theorem H.7 (ST classifier and its error on target)**.** *Starting with ERM solution, ST will lead to:*

(i) (Necessary condition) $h^t_{ST} = w_{\text{sp}}$ as $t \to \infty$, such that the target accuracy is 50% when the problem parameters $\gamma, \sigma_{\text{sp}}, d_{\text{sp}}$ satisfy:

$$\frac{\exp\left(-\sigma_0^2/50\right)}{\frac{4\sigma_0}{5\sqrt{2}} + \sqrt{\left(\frac{4\sigma_0}{5\sqrt{2}}\right)^2 + 4/\pi}} - \frac{\exp\left(-\sigma_0^2/50\right)}{\sigma_0} \leqslant \left(\frac{1}{\frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(\frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} - \frac{1}{\sigma_0}\right) \cdot \frac{\gamma^2}{\sigma_{\text{sp}}}\,, \tag{13}$$

where $\sigma_0 = \frac{\sigma_{\text{sp}}}{d_{\text{sp}} + \gamma^2}$.

(ii) (Sufficient condition) $h^t_{ST} = w_{\text{in}}$ as $t \to \infty$, such that the target accuracy is 100% when the problem parameters $\gamma, \sigma_{\text{sp}}, d_{\text{sp}}$ satisfy: $\sigma_{\text{sp}} \geqslant 1$ and $\gamma^2 \geqslant 2\sqrt{d_{\text{sp}}}\sigma_{\text{sp}}$.

Proof. The proof can be divided into two parts: (i) deriving closed-form expressions for updates on h^t_{ST} in terms of h^{t-1}_{ST} and (ii) obtaining conditions under which the component along w_{in} monotonically increases or decreases with t after re-normalizing the norm of updated h. For notation convenience, we denote h_{ST} with h in the rest of the proof.

Part-1. First, the loss of self-training with classifier $h := [h_{\text{in}}, h_{\text{sp}}]$ where $h_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$ and $h_{\text{sp}} \in \mathbb{R}^{d_{\text{sp}}}$ is given by:

$$\begin{aligned}
\mathcal{L}_{\text{st}}(h) &= \mathbb{E}_{\mathrm{P_T}(x)}\left[\ell(h^\top x, \operatorname{sgn}(h^\top x))\right] && (14)\\
&= \mathbb{E}_{\mathrm{P_T}(x)}\left[\exp\left(-\operatorname{sign}(h^\top x) \cdot (h^\top x)\right)\right] && (15)\\
&= \mathbb{E}_{\mathrm{P_T}(x)}\left[\exp\left(-\left|h^\top x\right|\right)\right] && (16)\\
&= \mathbb{E}_{\mathrm{P_T}(x)}\left[\exp\left(-\left|h_{\text{in}}^\top x_{\text{in}} + h_{\text{sp}}^\top x_{\text{sp}}\right|\right)\right] && (17)\\
&= \mathbb{E}_{y \sim U\{-1,1\}, z \sim \mathcal{N}(0,1)}\left[\exp\left(-\left|\gamma \cdot y \cdot h_{\text{in}}^\top w^\star \right.\right.\right. \\
&\qquad\quad \left.\left.\left. + \left[\sigma_{\text{in}}(\|h_{\text{in}}\|_2^2 - (h_{\text{in}}^T w^\star)^2) + \sigma_{\text{sp}} \cdot \|h_{\text{sp}}\|_2\right] \cdot z\right|\right)\right]\,. && (18)\\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[\exp\left(-\left|\gamma \cdot h_{\text{in}}^\top w^\star + \left[\sigma_{\text{in}}(\|h_{\text{in}}\|_2^2 - (h_{\text{in}}^T w^\star)^2) + \sigma_{\text{sp}} \cdot \|h_{\text{sp}}\|_2\right] \cdot z\right|\right)\right]\,, && (19)
\end{aligned}$$

where (17) to (18) is implied by simply replacing the definition of target distribution and (18) to (19) is implied by the symmetry of the function with respect to y and $-y$ due to the symmetry of the absolute function and Gaussian distribution.

For a classifier h^t, we denote $\mu_t = \gamma \cdot {h^t_{\text{in}}}^\top w^\star$ and $\sigma_t = \left[\sigma_{\text{in}}(\|h^t_{\text{in}}\|_2^2 - ({h^t_{\text{in}}}^T w^\star)^2) + \sigma_{\text{sp}} \cdot \left\|h^t_{\text{sp}}\right\|_2\right]$. With this notation, we can re-write the loss in (19) as $\mathcal{L}_{\text{st}}(h^t) = \mathbb{E}_{z \sim \mathcal{N}(0, \sigma_t^2)}\left[\exp\left(-\left|\mu_t + z\right|\right)\right]$.

Now we derive a closed-form expression of $\mathcal{L}_{\text{st}}(h^t)$ in Lemma J.10:

$$\mathcal{L}_{\text{st}}(h^t) = \frac{1}{2}\left(\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\right). \tag{20}$$

Define:

$$\alpha_1(\mu_t, \sigma_t) = -\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right), \tag{21}$$

$$\begin{aligned}\alpha_2(\mu_t, \sigma_t) = \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) \\ - \frac{2\sqrt{2}}{\sigma_t\sqrt{\pi}}\exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right).\end{aligned} \tag{22}$$

Let $\widetilde{h}^{t+1}$ denote the un-normalized gradient descent update at iterate $t+1$. We have:

$$\widetilde{h}^{t+1} = h^t - \eta \cdot \frac{\partial \mathcal{L}_{\text{st}}(h^t)}{\partial h}. \tag{23}$$

Now we will individually argue about the update of $\widetilde{h}^{t+1}$ along the first d_{in} dimensions and the last d_{sp} dimensions. First, we have:

$$\begin{aligned}\widetilde{h}^{t+1}_{\text{in}} &= h^t_{\text{in}} - \eta \cdot \frac{\partial \mathcal{L}_{\text{st}}(h^t)}{\partial h_{\text{in}}} \\ &= h^t_{\text{in}} - \frac{\eta}{2}\left(-\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\right. \\ &\qquad \left. + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\right) \cdot \gamma \cdot w^\star \\ &\qquad - \frac{\eta}{2}\left(\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\right. \\ &\qquad + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) \\ &\qquad \left. - \frac{2\sqrt{2}}{\sigma_t\sqrt{\pi}}\exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right)\right) \cdot (2h^t_{\text{in}} - 2({h^t_{\text{in}}}^\top w^\star)w^\star) \cdot \sigma_{\text{in}}^2 \\ &= h^t_{\text{in}} - \frac{\eta}{2} \cdot \alpha_1(\mu_t, \sigma_t) \cdot \gamma \cdot w^\star - \frac{\eta}{2} \cdot \alpha_2(\mu_t, \sigma_t) \cdot (2h^t_{\text{in}} - 2({h^t_{\text{in}}}^\top w^\star)w^\star) \cdot \sigma_{\text{in}}^2.\end{aligned} \tag{24}$$

Notice that the update of h^{t+1}_{in} is split into two components, one along $w^\star$ and the other along the orthogonal component $2h^t_{\text{in}} - 2({h^t_{\text{in}}}^\top w^\star)w^\star$. We will now argue that since at initialization, the component along $(I - w^\star {w^\star}^\top)$ is zero then it will remain zero. In particular, we have:

$${h^0_{\text{in}}}^\top (I - w^\star {w^\star}^\top) \propto {w^\star}^\top (I - w^\star {w^\star}^\top) = 0. \tag{25}$$

With (24), we can argue that if $(I - w^\star {w^\star}^\top)h^t_{\text{in}} = 0$, then $(I - w^\star {w^\star}^\top)\widetilde{h}^{t+1}_{\text{inv}} = 0$ implying that $(I - w^\star {w^\star}^\top)\widetilde{h}^t_{\text{in}} = 0$ for all $t > 0$. Hence, we have:

$$\begin{aligned} \widetilde{h}_{\text{inv}}^{t+1} &= h_{\text{in}}^t - \eta \cdot \frac{\partial \mathcal{L}_{\text{st}}(h^t)}{\partial h_{\text{in}}} \\ &= h_{\text{in}}^t - \frac{\eta}{2} \cdot \alpha_1(\mu_t, \sigma_t) \cdot \gamma \cdot w^\star \,. \end{aligned} \tag{26}$$

Second, we have the update $\widetilde{h}_{\text{sp}}^{t+1}$ given by:

$$\begin{aligned} \widetilde{h}_{\text{sp}}^{t+1} &= h_{\text{sp}}^t - \eta \cdot \frac{\partial \mathcal{L}_{\text{st}}(h^t)}{\partial h_{\text{sp}}} \\ &= h_{\text{sp}}^t - \frac{\eta}{2} \left(\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \operatorname{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) \right. \\ &\qquad \left. + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \operatorname{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) - \frac{2\sqrt{2}}{\sigma_t\sqrt{\pi}} \exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right) \right) \cdot h_{\text{sp}}^t \cdot \sigma_{\text{sp}}^2 \\ &= h_{\text{sp}}^t - \frac{\eta}{2} \cdot \alpha_2(\mu_t, \sigma_t) \cdot h_{\text{sp}}^t \cdot \sigma_{\text{sp}}^2 \,. \end{aligned} \tag{27}$$

Re-writing the expressions (26) and (27) for the update of $\widetilde{h}^{t+1}$, we have:

$$\widetilde{h}_{\text{in}}^{t+1} = h_{\text{in}}^t (1 - \frac{\eta}{2} \cdot \alpha_1(\mu_t, \sigma_t) \cdot \gamma^2/\mu_t) \,. \tag{28}$$

$$\widetilde{h}_{\text{sp}}^{t+1} = h_{\text{sp}}^t (1 - \frac{\eta}{2} \cdot \alpha_2(\mu_t, \sigma_t) \cdot \sigma_{\text{sp}}^2) \,. \tag{29}$$

Here, we replace $h_{\text{sp}}^t = \mu_t \cdot w^\star/\gamma$ in (26) to get (28). Updates in (28) and (29) show that $\widetilde{h}_{\text{inv}}^{t+1}$ remains in the direction of h_{in}^t and $\widetilde{h}_{\text{sp}}^{t+1}$ remains in the direction of h_{sp}^t.

Part-2. Now we will derive conditions under which h_{in}^t and h_{sp}^t will show monotonic behavior for necessary and sufficient conditions. We will first argue the condition under which ST will provably fail and converge to a classifier with a random target performance. For this, at every t, if we have:

$$\frac{\left\| \widetilde{h}_{\text{sp}}^{t+1} \right\|_2}{\left\| \widetilde{h}^{t+1} \right\|_2} > \left\| h_{\text{sp}}^t \right\|_2 \,, \tag{30}$$

then we can argue that as $t \to \infty$, we have $\left\| h_{\text{sp}}^t \right\|_2 = 1$ and hence, the ST classifier will have random target performance. Thus, we will focus on conditions, under which the norm on $\left\| h_{\text{sp}}^t \right\|_2$ increases with t. Re-writing (30), we have:

$$\left\| \widetilde{h}_{\text{sp}}^{t+1} \right\|_2 > \left\| \widetilde{h}^{t+1} \right\|_2 \cdot \left\| h_{\text{sp}}^t \right\|_2 \tag{31}$$

$$\left\| \widetilde{h}_{\text{sp}}^{t+1} \right\|_2 > \left(\left\| \widetilde{h}_{\text{sp}}^{t+1} \right\|_2 + \left\| \widetilde{h}_{\text{in}}^{t+1} \right\|_2 \right) \cdot \left\| h_{\text{sp}}^t \right\|_2 \tag{32}$$

$$\left\| \widetilde{h}_{\text{sp}}^{t+1} \right\|_2 \cdot \left(1 - \left\| h_{\text{sp}}^t \right\|_2 \right) > \left\| \widetilde{h}_{\text{in}}^{t+1} \right\|_2 \cdot \left\| h_{\text{sp}}^t \right\|_2 \tag{33}$$

$$\frac{\left\| \widetilde{h}_{\text{sp}}^{t+1} \right\|_2}{\left\| h_{\text{sp}}^t \right\|_2} > \frac{\left\| \widetilde{h}_{\text{in}}^{t+1} \right\|_2}{\left\| h_{\text{in}}^t \right\|_2} \,. \tag{34}$$

Plugging in (28) and (29) into (34), we get:

$$\left| 1 - \frac{\eta}{2} \cdot \alpha_2(\mu_t, \sigma_t) \cdot \sigma_{\text{sp}}^2 \right| > \left| 1 - \frac{\eta}{2} \cdot \alpha_1(\mu_t, \sigma_t) \cdot \gamma^2/\mu_t \right| \,. \tag{35}$$

For small enough η, we have the necessary condition for the failure of ST as:

$$\alpha_2(\mu_t, \sigma_t) \cdot \sigma_{\mathrm{sp}}^2 < \alpha_1(\mu_t, \sigma_t) \cdot \gamma^2/\mu_t \,. \tag{36}$$

Now we show in Lemma H.9 and Lemma H.8 that if the conditions assumed in the theorem continue to hold, then we can success and failure respectively.

□

Lemma H.8 (Necessary conditions for ST)**.** *Define α_1 and α_2 as in (21) and (22) respectively. Assume that $\frac{\partial}{\partial \mu}\alpha_2(\mu, \sigma) \geqslant 0$ for all $\mu \in [0, \mu_0]$. If $\sigma_{\mathrm{sp}} \geqslant 1$, $\mu_0 \leqslant \frac{\sigma_0^2}{5}$, and*

$$\frac{\exp\left(-\sigma_0^2/50\right)}{\frac{4\sigma_0}{5\sqrt{2}} + \sqrt{\left(\frac{4\sigma_0}{5\sqrt{2}}\right)^2 + 4/\pi}} - \frac{\exp\left(-\sigma_0^2/50\right)}{\sigma_0} \leqslant \left(\frac{1}{\frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(\frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} - \frac{1}{\sigma_0}\right) \cdot \frac{\gamma^2}{\sigma_{\mathrm{sp}}} \,, \tag{37}$$

then we have for all t:

$$\alpha_2(\mu_t, \sigma_t) \cdot \frac{\sigma_{\mathrm{sp}}^2 \cdot \mu_t}{\gamma^2} \leqslant \alpha_1(\mu_t, \sigma_t) \,. \tag{38}$$

Proof. We first recall that μ_t decreases and σ_t increases as (38) continues to hold true. We perform Taylor's expansion of $\alpha_1(\mu_t, \sigma_t)$ at $\mu_t = 0$. We have:

$$\alpha_1(\mu_t, \sigma_t) = \alpha_1(0, \sigma_t) + \left[\frac{\partial}{\partial \mu_t}\alpha_1(\mu_t, \sigma_t)\right]_{\mu_t = 0} \cdot \mu_t + \left[\frac{\partial^2}{\partial \mu_t^2}\alpha_1(\mu_t, \sigma_t)\right]_{\mu_t = \epsilon} \cdot \frac{\epsilon^2}{2} \,, \tag{39}$$

for some $\epsilon \in [0, \mu_t)$. Notice that $\left[\frac{\partial}{\partial \mu_t}\alpha_1(\mu_t, \sigma_t)\right]_{\mu_t = 0} = \alpha_2(0, \sigma_t)$. By assumption, we have $\left[\frac{\partial^2}{\partial \mu_t^2}.\alpha_1(\mu_t, \sigma_t)\right]_{\mu_t = \epsilon} \geqslant 0$. This implies that α_2 is increasing in μ in the interval $[0, \mu_0]$ and hence, the necessary condition reduces to the following:

$$\alpha_2(\mu_0, \sigma_0) \cdot \frac{\sigma_{\mathrm{sp}}^2}{\gamma^2} \leqslant \alpha_2(0, \sigma_0) \,. \tag{40}$$

We now use Lemma J.1 to obtain an upper bound on LHS and lower bound on RHS. In particular, we get:

$$\alpha_2(\mu_0, \sigma_0) \leqslant \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_0^2/(2 \cdot \sigma_0^2)\right)}{-\frac{\mu_0}{\sqrt{2}\sigma_0} + \frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_0}{\sqrt{2}\sigma_0} + \frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} \tag{41}$$

$$+ \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_0^2/(2 \cdot \sigma_0^2)\right)}{\frac{\mu_0}{\sqrt{2}\sigma_0} + \frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(\frac{\mu_0}{\sqrt{2}\sigma_0} + \frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} - \frac{2\sqrt{2}}{\sigma_0\sqrt{\pi}}\exp\left(-\frac{\mu_0^2}{2\sigma_0^2}\right) . \tag{42}$$

when $\mu_0 \leqslant \sigma_0^2/5$, we have:

$$\alpha_2(\mu_0, \sigma_0) \leqslant \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\sigma_0^2/50\right)}{\frac{4\sigma_0}{5\sqrt{2}} + \sqrt{\left(\frac{4\sigma_0}{5\sqrt{2}}\right)^2 + 4/\pi}} + \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\sigma_0^2/50\right)}{\frac{6\sigma_0}{5\sqrt{2}} + \sqrt{\left(\frac{6\sigma_0}{5\sqrt{2}}\right)^2 + 4/\pi}} - \frac{2\sqrt{2}\exp\left(-\sigma_0^2/50\right)}{\sigma_0\sqrt{\pi}} \,, \tag{43}$$

Similarly, we have:

$$\alpha_1(0, \sigma_0) \geqslant \frac{2}{\sqrt{\pi}} \frac{1}{\frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(\frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} + \frac{2}{\sqrt{\pi}} \frac{1}{\frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(\frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} - \frac{2\sqrt{2}}{\sigma_0\sqrt{\pi}} \,, \tag{44}$$

Thus, if we have:

$$\frac{\exp\left(-\sigma_0^2/50\right)}{\frac{4\sigma_0}{5\sqrt{2}} + \sqrt{\left(\frac{4\sigma_0}{5\sqrt{2}}\right)^2 + 4/\pi}} - \frac{\exp\left(-\sigma_0^2/50\right)}{\sigma_0} \leqslant \left(\frac{1}{\frac{\sigma_0}{\sqrt{2}} + \sqrt{\left(\frac{\sigma_0}{\sqrt{2}}\right)^2 + 4/\pi}} - \frac{1}{\sigma_0}\right) \cdot \frac{\gamma^2}{\sigma_{\text{sp}}}, \tag{45}$$

then $\alpha_2(\mu_t, \sigma_t) \cdot \frac{\sigma_{\text{sp}}^2 \cdot \mu_t}{\gamma^2} \leqslant \alpha_1(\mu_t, \sigma_t)$ will continue to hold for all t.

□

Lemma H.9 (Sufficiency conditions for ST)**.** *Define α_1 and α_2 as in* (21) *and* (22) *respectively. If $\sigma_{\text{sp}} \geqslant 1$ and $\mu_0 \geqslant 2\sigma_0^2$, then we have for all t:*

$$\alpha_2(\mu_t, \sigma_t) \cdot \frac{\sigma_{\text{sp}}^2 \cdot \mu_t}{\gamma^2} \geqslant \alpha_1(\mu_t, \sigma_t) \,. \tag{46}$$

Proof. We first recall that μ_t increases and σ_t decreases as (46) continues to hold true. First, we use Lemma J.1, to obtain an upper bound on $\alpha_1(\mu_t, \alpha_t)$:

$$\begin{aligned}
\alpha_1(\mu_t, \alpha_t) &= -\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \operatorname{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \operatorname{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) && (47)\\
&\leqslant -\frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} + \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}}, && (48)\\
&\leqslant 0\,, && (49)
\end{aligned}$$

whenever $\left(-\frac{\mu_t}{2\sigma_t} + \frac{\sigma_t}{2}\right)^2 + 1 \leqslant \frac{\mu_t^2}{\sigma_t^2}$. Simplifying this further, we get $\frac{\sigma_t^2}{4} + 1 \leqslant \frac{3\mu_t^2}{4\sigma_t^2}$. Moreover, since μ_t is increasing and σ_t is decreasing, if we have $\frac{\sigma_0^2}{4} + 1 \leqslant \frac{3\mu_0^2}{4\sigma_0^2}$ then $\alpha(\mu_t, \sigma_t) \leqslant 0$ for all iterations.

Now, we will show that under the assumed conditions $\alpha_2(\mu_t, \sigma_t)$ is lower bounded by zero. In particular, we use Lemma J.1, to obtain an lower bound on $\alpha_2(\mu_t, \alpha_t)$. Recall,

$$\begin{aligned}
\alpha_2(\mu_t, \sigma_t) &= \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \operatorname{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \operatorname{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) && (50)\\
&\qquad - \frac{2\sqrt{2}}{\sigma_t\sqrt{\pi}} \exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right). && (51)\\
&> \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \operatorname{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) - \frac{2\sqrt{2}}{\sigma_t\sqrt{\pi}} \exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right) && (52)\\
&> \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} - \frac{2\sqrt{2}}{\sigma_t\sqrt{\pi}} \exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right) && (53)\\
&> \frac{2}{\sqrt{\pi}} \cdot \exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right) \left(\frac{1}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} - \frac{\sqrt{2}}{\sigma_t}\right) && (54)\\
&> \frac{2}{\sqrt{\pi}} \cdot \exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right) \cdot \frac{\sqrt{2}}{\sigma_t} \left(\frac{1}{-\frac{\mu_t}{\sigma_t^2} + 1 + \sqrt{\left(-\frac{\mu_t}{\sigma_t^2} + 1\right)^2 + \frac{2\sqrt{2}}{\sigma_t}}} - 1\right) && (55)\\
&> 0\,, && (56)
\end{aligned}$$

whenever $\mu_t \geqslant 2\sigma_t^2$ and $\sigma_{\text{sp}} \geqslant 1$ which further implies that since μ_t is increasing and σ_t is decreasing, if we have $\mu_0 \geqslant 2\sigma_0^2$ then $\alpha_2(\mu_t, \sigma_t)$ remains positive.

□

H.6. Analysis of CL

H.6.1. Proof of Proposition H.2

For convenience, we first restate the Proposition H.2 which gives us a closed form solution for (6) when $\rho = 0$. Then, we provide the proof, focusing first on the case of $k = 1$, and then showing that extension to $k > 1$ is straightforward and renders the final form in the proposition that follows.

Proposition H.10 (Barlow Twins solution)**.** *The solution for* (6) *is* $U_k^\top \Sigma_{\mathsf{A}}^{-1/2}$ *where* U_k *are the top* k *eigenvectors of* $\Sigma_{\mathsf{A}}^{-1/2}\,\widetilde{\Sigma}\,\Sigma_{\mathsf{A}}^{-1/2}$. *Here,* $\Sigma_{\mathsf{A}} := \mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[aa^\top]$ *is the covariance over augmentations, and* $\widetilde{\Sigma} := \mathbb{E}_{x\sim \mathrm{P}_{\mathsf{U}}}[\widetilde{a}(x)\widetilde{a}(x)^\top]$ *is the covariance matrix of mean augmentations* $\widetilde{a}(x) := \mathbb{E}_{\mathrm{P}_{\mathsf{A}}(a|x)}[a]$.

Proof. We will use $\phi(x)$ to denote $\phi^\top x$ where $\phi \in \mathbb{R}^d$. Throughout the proof, we use a to denote augmentation and x to denote the input. We will use $\mathrm{P}_{\mathsf{A}}(a \mid x)$ as the probability measure over the space of augmentations $\mathcal{A}$, given some input $x \in \mathcal{X}$ (with corresponding density) $p_{\mathsf{A}}(\cdot \mid x)$. Next, we use $p_{\mathsf{A}}(\cdot)$ to denote the density associate with the marginal probability measure over augmentations: $\mathrm{P}_{\mathsf{A}} = \int_{\mathcal{X}} \mathrm{P}_{\mathsf{A}}(a \mid x)\mathrm{dP}_{\mathsf{U}}$. Finally, the joint distribution over positive pairs $A_+(a_1, a_2) = \int_{\mathcal{X}} \mathrm{P}_{\mathsf{A}}(a_1 \mid x)\mathrm{P}_{\mathsf{A}}(a_2 \mid x)\mathrm{dP}_{\mathsf{U}}$, gives us the positive pair graph over augmentations.

Before we solve the optimization problem in (6) for $\Phi \in \mathbb{R}^{k\times d}$ for any general k, let us first consider the case where $k = 1$, *i.e.* we only want to find a single linear projection ϕ. The constraint $\rho = 0$, transfers onto ϕ in the following way:

$$\mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[\phi(a)^2] = 1 \quad \equiv \quad \phi^\top \Sigma_A \phi = 1 \tag{57}$$

Under the above constraint we want to minimize the invariance loss, which according to Lemma J.2 is given by $2 \cdot \int_{\mathcal{A}} \phi(a) L(\phi)(a)\ \mathrm{dP}_{\mathsf{A}}$, where $L(\phi)(\cdot)$ is the following linear operator.

$$L(\phi)(a) = \phi(a) - \int_{\mathcal{A}} \frac{A_+(a, a')}{p_{\mathsf{A}}(a)} \cdot \phi(a')\ \mathrm{d}a'. \tag{58}$$

Based on the definition of the operator, we can reformulate the constrained optimization for contrastive pretraining as:

$$\begin{aligned}
&\mathop{\arg\min}_{\phi:\phi^\top\Sigma_A\phi=1} \int_{\mathcal{A}} \phi(a)\cdot L(\phi)(a)\ \mathrm{dP}_{\mathsf{A}} && (59)\\
\implies &\mathop{\arg\min}_{\phi:\phi^\top\Sigma_A\phi=1}\ \mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[\phi(a)^2] - \int_{\mathcal{A}}\int_{\mathcal{A}} \phi(a)\cdot\phi(a')\cdot A_+(a,a')\ \mathrm{d}a\mathrm{d}a' && (60)\\
\implies &\mathop{\arg\min}_{\phi:\phi^\top\Sigma_A\phi=1}\ \mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[\phi(a)^2] - \int_{\mathcal{X}}\int_{\mathcal{A}}\int_{\mathcal{A}} p_{\mathsf{A}}(a\mid x)p_{\mathsf{A}}(a'\mid x)\cdot\phi(a)\phi(a')\ \mathrm{dP}_{\mathsf{U}} && (61)\\
\implies &\mathop{\arg\min}_{\phi:\phi^\top\Sigma_A\phi=1}\ \mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[\phi(a)^2] - \int_{\mathcal{X}} [\widetilde{\phi}(x)]^2\ \mathrm{dP}_{\mathsf{U}}, && (62)
\end{aligned}$$

where $\widetilde{\phi}(x) = \mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}(\cdot|x)}\phi(x) = \mathbb{E}_{c\sim \mathrm{Unif}[0,1]^d}[\phi^\top(c \odot x)]$. Note that,

$$\begin{aligned}
\widetilde{\phi}(x)^2 &= \left(\mathbb{E}_{c\sim \mathrm{Unif}[0,1]^d}[\phi^\top(c\odot x)]\right)^2 && (63)\\
&= \phi^\top(\mathbb{E}_{c\sim \mathrm{Unif}[0,1]^d}[c\odot x])(\mathbb{E}_{c\sim \mathrm{Unif}[0,1]^d}[c\odot x])^\top\phi && (64)\\
\implies \int_{\mathcal{X}} [\widetilde{\phi}(x)]^2\ \mathrm{dP}_{\mathsf{U}} &= \phi^\top\widetilde{\Sigma}\phi && (65)
\end{aligned}$$

Further, since $\mathbb{E}_{a\sim \mathrm{P}_{\mathsf{A}}}[\phi(a)^2] = \phi^\top\Sigma\phi$ we can now rewrite our main optimization problem for $k = 1$ as:

$$\underset{\phi:\phi^\top\Sigma_A\phi=1}{\arg\min} \quad \phi^\top\Sigma_A\phi - \phi^\top\widetilde{\Sigma}\phi \tag{66}$$

$$= \underset{\phi:\phi^\top\Sigma_A\phi=1}{\arg\max} \quad \phi^\top\widetilde{\Sigma}\phi \tag{67}$$

Recall that in our setup both $\widetilde{\Sigma}$ and Σ_A are positive definite and invertible matrices. To solve the above problem, let's consider a re-parameterization: $\phi' = \Sigma_A^{1/2}\phi$, thus $\phi^\top\Sigma_A\phi = 1$, is equivalent to the constraint $\|\phi'\|_2^2 = 1$. Based on this re-parameterization we are now solving:

$$\underset{\|\phi'\|_2^2=1}{\arg\max} \quad \phi'^\top\Sigma_A^{-1/2} \cdot \widetilde{\Sigma} \cdot \Sigma_A^{-1/2}\phi', \tag{68}$$

which is nothing but the top eigenvector for $\Sigma_A^{-1/2} \cdot \widetilde{\Sigma} \cdot \Sigma_A^{-1/2}$.

Now, to extend the above argument from $k = 1$ to $k > 1$, we need to care of one additional form of constraint in the form of feature diversity: $\phi_i^\top\Sigma_A\phi_j = 0$ when $i \neq j$. But, we can easily redo the reformulations above and arrive at the following optimization problem:

$$\underset{\substack{\|\phi_i'\|_2^2 = 1, \ \forall i \\ \phi_i'^\top\phi_j' = 0, \ \forall i \neq j}}{\arg\max} \quad \left[\phi_1', \phi_2', \ldots, \phi_k'\right]^\top \Sigma_A^{-1/2} \cdot \widetilde{\Sigma} \cdot \Sigma_A^{-1/2} \left[\phi_1', \phi_2', \ldots, \phi_k'\right], \tag{69}$$

where $\phi_i' = \Sigma_A^{1/2}\phi_i$. The above is nothing but the top k eigenvectors for the matrix $\Sigma_A^{-1/2} \cdot \widetilde{\Sigma} \cdot \Sigma_A^{-1/2}$. This completes the proof of Proposition H.10. □

H.6.2. Analysis with $\rho > 0$ in Contrastive Pretraining Objective (6)

In (6) we considered the strict version of the optimization problem where $\rho = 0$. Here, we will consider the following optimization problem that we optimize for our experiments in the simplified setup:

$$\mathcal{L}_{\mathrm{cl}}(\Phi, \kappa) \; := \; \mathbb{E}_{x\sim \mathrm{P_U}}\mathbb{E}_{a_1,a_2\sim \mathrm{P_A}(\cdot|x)} \; \|\Phi(a_1) - \Phi(a_2)\|_2^2 + \kappa \cdot \left\|\mathbb{E}_{a\sim \mathrm{P_A}}\left[\Phi(a)\Phi(a)^\top\right] - \mathbf{I}_k\right\|_F^2, \tag{70}$$

where $\kappa > 0$ is some finite constant (note that every ρ corresponds to some κ and particularly $\rho = 0$, corresponds to $\kappa = \infty$). Let $\Phi^\star$ be the solution for (6) with $\rho = 0$, *i.e.* the solution described in Proposition H.2. Now, we will show that in practice we can provably recover something close to $\Phi^\star$ when κ is large enough.

Theorem H.11 (Solution for (70) is approximately equal to $\Phi^\star$)**.** *If $\widehat{\Phi}$ is some solution that achieves low values of the objective $\mathcal{L}_{\mathrm{cl}}(\Phi, \kappa)$ in (70),* i.e.*, $\mathcal{L}_{\mathrm{cl}}(\widehat{\Phi}, \kappa) \leqslant \epsilon$, then there exists matrix $W \in \mathbb{R}^{k\times k}$ such that:*

$$\mathbb{E}_{a\sim \mathrm{P_A}}\|W \cdot \Phi^\star(a) - \widehat{\Phi}(a)\|_2^2 \leqslant \frac{k\epsilon}{2\gamma_{k+1}},$$

$$\textit{where,} \quad \gamma_{k+1} \geqslant \frac{2\gamma_1^2}{k\epsilon} \cdot \left(1 - \sqrt{\frac{\epsilon}{\kappa}}\right) - \frac{\gamma_1}{k},$$

where γ_{k+1} is the the $k + 1^{th}$ eigenvalue for $\mathbf{I}_d - \Sigma_A^{-1/2}\, \widetilde{\Sigma}\, \Sigma_A^{-1/2}$. Here, $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_d$.

Proof. Since we know that $\mathcal{L}_{\mathrm{cl}}(\widehat{\Phi}, \kappa) \leqslant \epsilon$, we can individually bound the invariance loss and the regularization term:

$$\mathbb{E}_{x\sim \mathrm{P_U}}\mathbb{E}_{a_1,a_2\sim \mathrm{P_A}(\cdot|x)} \; \|\widehat{\Phi}(a_1) - \widehat{\Phi}(a_2)\|_2^2 \leqslant \epsilon \tag{71}$$

$$\left\|\mathbb{E}_{a\sim \mathrm{P_A}}\left[\widehat{\Phi}(a)\widehat{\Phi}(a)^\top\right] - \mathbf{I}_k\right\|_F^2 \leqslant \frac{\epsilon}{\kappa} \tag{72}$$

Thus,

$$\forall i \in [k]: \quad 1 - \sqrt{\frac{\epsilon}{\kappa}} \leqslant \widehat{\phi}_i^\top \ \Sigma_A \ \widehat{\phi}_i \leqslant 1 + \sqrt{\frac{\epsilon}{\kappa}} \tag{73}$$

$$\forall i \in [k]: \quad \mathbb{E}_{x\sim \mathrm{P_U}} \mathbb{E}_{a_1,a_2 \sim \mathrm{P_A}(\cdot|x)} (\widehat{\phi}_i^\top a_1 - \widehat{\phi}_i^\top a_2)^2 \leqslant \epsilon \tag{74}$$

Let $\phi_1^\star, \phi_2^\star, \phi_3^\star, \ldots, \phi_d^\star$ be the solution returned by the analytical solution for $\rho = 0$, *i.e*. the solution in Proposition H.2. Now, since $\Phi^\star$ would span $\mathbb{R}^d$ when Σ_A is full rank, we can denote:

$$\widehat{\phi}_i = \sum_{j=1}^{d} \eta_i^{(j)} \phi_j^\star \tag{75}$$

Now from Lemma J.2, the invariance loss for $\widehat{\phi}_i$ can be written using the operator $L(\phi)(a) = \phi(a) - \int_{\mathcal{A}} \frac{A_+(a,a')}{p_{\mathrm{A}}(a)} \phi(a') \ \mathrm{d}a'$:

$$\begin{aligned}
\text{Invariance Loss}(\widehat{\phi}_i) &\coloneqq \mathbb{E}_{x\sim \mathrm{P_U}} \mathbb{E}_{a_1,a_2 \sim \mathrm{P_A}(\cdot|x)} (\widehat{\phi}_i^\top a_1 - \widehat{\phi}_i^\top a_2)^2 && (76)\\
&= 2 \cdot \mathbb{E}_{a\sim \mathrm{P_A}} [\widehat{\phi}_i(a) L(\widehat{\phi}_i)(a)] && (77)\\
&= 2 \cdot \mathbb{E}_{a\sim \mathrm{P_A}} \left[\left(\sum_{j=1}^{d} \eta_i^{(j)} \phi_i^\star \right) L \left(\sum_{j=1}^{d} \eta_i^{(j)} \phi_j^\star \right)(a) \right] && (78)\\
&= 2 \cdot \mathbb{E}_{a\sim \mathrm{P_A}} \left[\left(\sum_{j=1}^{d} \eta_i^{(j)} \phi_j^\star \right) \left(\sum_{j=1}^{d} \eta_i^{(j)} L(\phi_j^\star)(a) \right) \right] && (79)\\
&= 2 \cdot \sum_{j=1}^{d} \left(\eta_i^{(j)}\right)^2 \mathbb{E}_{a\sim \mathrm{P_A}} \left[\phi_j^\star(a) L(\phi_j^\star)(a)\right] && (80)\\
&\quad + 2 \cdot \sum_{m=1,n=1,m\neq n}^{d} \eta_i^{(m)} \eta_i^{(n)} \mathbb{E}_{a\sim \mathrm{P_A}} \left[\phi_m^\star(a) L(\phi_n^\star)(a)\right] && (81)
\end{aligned}$$

Since, $\phi_i^\star(\cdot)$ are eigenfunctions of the operator L (HaoChen & Ma, 2022), we can conclude that:

$$\sum_{m=1,n=1,m\neq n}^{d} \eta_i^{(m)} \eta_i^{(n)} \mathbb{E}_{a\sim \mathrm{P_A}} \left[\phi_m^\star(a) L(\phi_n^\star)(a)\right] = 0,$$

and if $\gamma_1 \leqslant \gamma_2 \leqslant \gamma_3 \ldots \leqslant \gamma_d$ are the eigenvalues for $\phi_1^\star, \phi_2^\star, \phi_3^\star, \ldots, \phi_d^\star$ under the decomposition of $L(\phi)(\cdot)$ then:

$$\mathbb{E}_{x\sim \mathrm{P_U}} \mathbb{E}_{a_1,a_2 \sim \mathrm{P_A}(\cdot|x)} (\widehat{\phi}_i^\top a_1 - \widehat{\phi}_i^\top a_2)^2 = 2 \cdot \sum_{j=1}^{d} \gamma_j \left(\eta_i^{(j)}\right)^2 \tag{82}$$

Recall, we are also aware of a condition on the regularization term: $1 - \sqrt{\frac{\epsilon}{\kappa}} \leqslant \widehat{\phi}_i^\top \ \Sigma_A \ \widehat{\phi}_i \leqslant 1 + \sqrt{\frac{\epsilon}{\kappa}}$.

$$\widehat{\phi}_i^\top \ \Sigma_A \ \widehat{\phi}_i = \left(\sum_{j=1}^{d} \eta_i^{(j)} \phi_j^\star \right)^\top \ \Sigma_A \ \left(\sum_{j=1}^{d} \eta_i^{(j)} \phi_j^\star \right) = \sum_{j=1}^{d} \left(\eta_i^{(j)}\right)^2 \tag{83}$$

$$\implies 1 - \sqrt{\frac{\epsilon}{\kappa}} \leqslant \sum_{j=1}^{d} \left(\eta_i^{(j)}\right)^2 \leqslant 1 + \sqrt{\frac{\epsilon}{\kappa}} \ \ \forall i. \tag{84}$$

In order to show that the projection of $\widehat{\phi}_i$ on Φ^* is significant, we need to argue that the term $\sum_{j=k+1}^{d} \left(\eta_i^{(j)}\right)^2$ is small. The argument for this begins with the condition on invariance loss, and the fact that $\gamma_1 \leqslant \gamma_2 \leqslant \ldots \leqslant \gamma_k \leqslant \gamma_{k+1} \leqslant \ldots \leqslant \gamma_d$:

$$\frac{\epsilon}{2} \geqslant \sum_{j=k+1}^{d} \left(\eta_i^{(j)}\right)^2 \gamma_j \geqslant \gamma_{k+1} \cdot \left(\sum_{j=k+1}^{d} \left(\eta_i^{(j)}\right)^2\right) \tag{85}$$

$$\implies \sum_{j=k+1}^{d} \left(\eta_i^{(j)}\right)^2 \leqslant \frac{\epsilon}{2\gamma_{k+1}} \tag{86}$$

Extending the above result $\forall i$ by simply adding the bounds completes the claim of our first result in Theorem H.11. Next, we will lower bound the eigenvalue γ_{k+1}. Recall that, $\sum_{j=1}^{k} \left(\eta_i^{(j)}\right)^2 \geqslant 1 - \sqrt{\frac{\epsilon}{\kappa}} - \frac{\epsilon}{2\gamma_{k+1}}$. Thus,

$$\gamma_1 \cdot \left(1 - \sqrt{\frac{\epsilon}{\kappa}} - \frac{\epsilon}{2\gamma_{k+1}}\right) \leqslant \sum_{j=1}^{k} \gamma_j \left(\eta_i^{(j)}\right)^2 \leqslant k\gamma_{k+1} \cdot \frac{\epsilon}{2\gamma_1} \tag{87}$$

We assume that all eigenvalues are strictly positive, which is true under our augmentation distribution. Given, $\gamma_{k+1} \geqslant \gamma_1$, we can rearrange the above to get:

$$\gamma_{k+1} \geqslant \frac{2\gamma_1^2}{k\epsilon} \cdot \left(1 - \sqrt{\frac{\epsilon}{\kappa}}\right) - \frac{\gamma_1}{k} \tag{88}$$

This completes the claim of our second result in Theorem H.11. □

H.6.3. Proof of Theorem H.3

Recall the definition of $w_{\text{in}} := [w^\star, 0, \ldots, 0]$ and $w_{\text{sp}} := [0, \ldots 0, w']$ where $w' = \mathbf{1}_{d_{\text{sp}}}/\sqrt{d_{\text{sp}}}$. Let us now define u_1, u_2 as the top two eigenvectors of Σ_A with eigenvalues $\lambda_1, \lambda_2 > 0$, (note that in our problem setup both Σ_A and $\widetilde{\Sigma}$ are full rank positive definite matrices), and $\tau := \sqrt{\lambda_1/\lambda_2}$. Next we define α as the angle between u_1 and w_{in}, *i.e.*, $\cos(\alpha) = u_1^\top w_{\text{in}}$. Based on the definitions of α and τ, both of which are fully determined by the eigen decomposition of the post-augmentation feature covariance matrix Σ_A we can re-write Theorem H.3 formally as:

Theorem H.12 (Formal; CL recovers both invariant w_{in} and spurious w_{sp} but amplifies w_{in})**.** *For $w^\star = {}^{\mathbf{1}_{d_{\text{in}}}}/\sqrt{d_{\text{in}}}$, the CL solution Φ_{cl}=$[\phi_1, \phi_2, ..., \phi_k]$ satisfies $\phi_j^\top w_{\text{in}} = \phi_j^\top w_{\text{sp}} = 0 \ \forall j \geqslant 3$. For τ, α as defined above, the solution for ϕ_1, ϕ_2 is:*

$$\begin{bmatrix} w^\star \cdot \cot(\alpha)/\tau, & w^\star \\ w' \cdot 1/\tau, & w' \cdot \cot(\alpha) \end{bmatrix} \cdot \begin{bmatrix} \cos\theta, & \sin\theta \\ \sin\theta, & -\cos\theta \end{bmatrix},$$

where $0 \leqslant \alpha, \theta \leqslant \pi/2$. Now, if we redefine $\phi_1 = c_1 w_{\text{in}} + c_3 w_{\text{sp}}$ and $\phi_2 = c_2 w_{\text{in}} + c_4 w_{\text{sp}}$, then $\forall \gamma, d_{\text{in}}, d_{\text{sp}}, \sigma_{\text{in}}$ satisfying $\gamma/\sqrt{d_{\text{sp}}} < p_0$ and $\sigma_{\text{in}}/\gamma < p_1$, we can show that $\exists \sigma_{\text{sp}_1}, \sigma_{\text{sp}_2}$ such that when $\sigma_{\text{sp}_1} \leqslant \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$, the the signal along w_{in} is amplified, i.e.*:*

- *In ϕ_1, we have ${}^{5\gamma}/\sqrt{d_{\text{sp}}} \leqslant c_1/c_3 \leqslant {}^{20\gamma}/\sqrt{d_{\text{sp}}}$.*
- *In ϕ_2, for some small $\epsilon > 0$, we have $|c_2/c_4| \geqslant (1-\epsilon) \cdot \sqrt{d_{\text{sp}}}/\gamma$.*

Here, it is sufficient for the constants p_0, p_1 to satisfy $p_0, p_1 \ll 1$. For e.g.*, $p_0 = 0.15$, $p_1 = 0.5$ (satisfied by our problem parameters in Example 1) are sufficient for our arguments to hold.*

Proof. We will first show that the only components of interest are ϕ_1, ϕ_2. Then, we will prove conditions on the amplification of w_{in} over w_{sp} in ϕ_1, ϕ_2. Following is the proof overview:

I. From the derived closed form expressions for Σ_A and $\widetilde{\Sigma}$, show that the solution returned by solving the Barlow Twins objective depends on w_{in} and w_{sp} only through the first two components ϕ_1, ϕ_2, when $w^\star = \mathbf{1}_{d_{\text{in}}}/\sqrt{d_{\text{in}}}$.

II. For the components ϕ_1, ϕ_2, we will show that the dependence along w_{in} is amplified over that on w_{sp} when the target data sufficiently denoises the spurious feature.

Part-I:

We can divide the space $\mathbb{R}^d$ into two subspaces that are perpendicular to each other. The first subspace is $\mathcal{W} = \{b_1 \cdot w_{\text{in}} + b_2 \cdot w_{\text{sp}} : b_1, b_2 \in \mathbb{R}\}$, *i.e.* the rank 2 subspace spanned by w_{in} and w_{sp}. The second subspace is $\mathcal{W}_\perp$ where $\mathcal{W}_\perp = \{u \in \mathbb{R}^d : u^\top w_{\text{in}} = 0, u^\top w_{\text{sp}} = 0\}$. Then, from Lemma J.3 we can conclude that the matrix Σ_A can be written as:

$$\Sigma_A = \Sigma_{A_\mathcal{W}} + \Sigma_{A_{\mathcal{W}_\perp}} \tag{89}$$

$$\Sigma_{A_\mathcal{W}} = \frac{1}{4}\begin{bmatrix} \left(\gamma^2(1+{}^1\!/\!{}_{3d_{\text{in}}}) + {}^{\sigma_{\text{in}}^2}\!/\!{}_3(1-{}^1\!/\!{}_{d_{\text{in}}})\right)\cdot w^\star {w^\star}^\top, & \gamma\sqrt{d_{\text{sp}}}/2 \cdot w^\star {w'}^\top \\ \gamma\sqrt{d_{\text{sp}}}/2 \cdot w' {w^\star}^\top, & \left({}^{d_{\text{sp}}}\!/\!{}_2 + {}^4\!/\!{}_3 \cdot \sigma_{\text{sp}}^2 + {}^1\!/\!{}_6\right)\cdot w' {w'}^\top \end{bmatrix}, \tag{90}$$

where $\Sigma_{A_{\mathcal{W}_\perp}} := \mathbb{E}_{a\sim \mathrm{P_A}}\left[\Pi_{\mathcal{W}_\perp}(a)(\Pi_{\mathcal{W}_\perp}(a))^\top\right]$ is the covariance matrix in the null space of $\mathcal{W}$, *i.e.* the covariance matrix in the space of non-predictive (noise) features. Similarly we can define:

$$\widetilde{\Sigma} \;=\; \widetilde{\Sigma}_\mathcal{W} + \widetilde{\Sigma}_{\mathcal{W}_\perp} \tag{91}$$

$$\widetilde{\Sigma}_\mathcal{W} \;=\; \frac{1}{4}\begin{bmatrix} \gamma^2 \cdot w^\star {w^\star}^\top, & \gamma\sqrt{d_{\text{sp}}}/2 \cdot w^\star {w'}^\top \\ \gamma\sqrt{d_{\text{sp}}}/2 \cdot w' {w^\star}^\top, & \left({}^{d_{\text{sp}}}\!/\!{}_2 + {}^{\sigma_{\text{sp}}^2}\!/\!{}_2\right)\cdot w' {w'}^\top \end{bmatrix} \tag{92}$$

Here again $\widetilde{\Sigma}_{\mathcal{W}_\perp} := \mathbb{E}_{x\sim \mathrm{P_U}}\left[\Pi_{\mathcal{W}_\perp}(\mathbb{E}_{c\sim \mathrm{Unif}[0,1]^d}(c\odot x))(\Pi_{\mathcal{W}_\perp}(\mathbb{E}_{c\sim \mathrm{Unif}[0,1]^d}(c\odot x)))^\top\right]$ is the covariance matrix of mean augmentations after they are projected onto the null space of predictive features. The above decomposition also follows from result in Lemma J.3.

From Proposition H.2 we know that the closed form expression for the solution returned by optimizing the Barlow Twins objective in (6) is $U^\top \Sigma_A^{-1/2}$ where U are the top-k eigenvectors of:

$$\Sigma_A^{-1/2}\cdot\widetilde{\Sigma}\cdot\Sigma_A^{-1/2} \tag{93}$$

When $w^\star = \mathbf{1}_{d_{\text{in}}}/\sqrt{d_{\text{in}}}$, then $\Sigma_{A_{\mathcal{W}_\perp}} = \widetilde{\Sigma}_{\mathcal{W}_\perp} + \frac{1}{3}\cdot\mathrm{diag}(\widetilde{\Sigma}_{\mathcal{W}_\perp})$. Further, since $\mathrm{diag}(\widetilde{\Sigma}_{\mathcal{W}_\perp}) = p\cdot \mathbf{I}_d$ for some constant p, the eigenvectors of $\widetilde{\Sigma}_{\mathcal{W}_\perp}$ and $\Sigma_{A_{\mathcal{W}_\perp}}$ are exactly the same. Hence, when we consider the SVD of the expression $\Sigma_A^{-1/2}\widetilde{\Sigma}\Sigma_A^{-1/2}$, the matrices $\Sigma_{A_{\mathcal{W}_\perp}}$ and $\widetilde{\Sigma}_{\mathcal{W}_\perp}$ have no effect on the SVD components that lie along the span of the predictive features. In fact, we only need to consider two rank 2 matrices (first terms in (91), (89)) and only do the SVD of $\Sigma_{A_\mathcal{W}}^{-1/2}\cdot\widetilde{\Sigma}_\mathcal{W}\cdot\Sigma_{A_\mathcal{W}}^{-1/2}$.

There are only two eigenvectors of $\Sigma_{A_\mathcal{W}}^{-1/2}\cdot\widetilde{\Sigma}_\mathcal{W}\cdot\Sigma_{A_\mathcal{W}}^{-1/2}$. We use λ_1, λ_2 to denote the eigenvalues of $\Sigma_{A_\mathcal{W}}$, and $[\cos(\alpha)w^\star, \sin(\alpha)w']^\top$, $[\sin(\alpha)w^\star, -\cos(\alpha)w']^\top$ for the corresponding eigenvectors. Similarly, we use $\widetilde{\lambda}_1, \widetilde{\lambda}_2$ to denote the eigenvalues of $\widetilde{\Sigma}_\mathcal{W}$, and $[\cos(\beta)w^\star, \sin(\beta)w']^\top$, $[\sin(\beta)w^\star, -\cos(\beta)w']^\top$ for the corresponding eigenvectors. Let $\mathrm{SVD}_U(\cdot)$ denote the operation of obtaining the singular vectors of a matrix. Then, to compute the components of the final expression: $\mathrm{SVD}_U(\Sigma_A^{-1/2}\widetilde{\Sigma}\Sigma_A^{-1/2})^\top\Sigma_A^{-1/2}$ that lies along the span of predictive features (in $\mathcal{W}$), we need only look at the decomposition of the following matrix:

$$\begin{bmatrix}\cos\theta & , \sin(\theta) \\ \sin\theta & , -\cos(\theta)\end{bmatrix} = \mathrm{SVD}_U\left(\begin{bmatrix} 1/\sqrt{\lambda_1}, & 0 \\ 0, & 1/\sqrt{\lambda_2}\end{bmatrix}\cdot\begin{bmatrix}\cos(\alpha-\beta), & \sin(\alpha-\beta) \\ \sin(\alpha-\beta), & -\cos(\alpha-\beta)\end{bmatrix}\cdot\begin{bmatrix}\sqrt{\widetilde{\lambda}_1}, & 0 \\ 0, & \sqrt{\widetilde{\lambda}_2}\end{bmatrix}\right) \tag{94}$$

Based on the above definitions of $\theta, \alpha, \lambda_1, \lambda_2$, we can then formulate ϕ_1 and ϕ_2 in the following way:

$$[\phi_1, \phi_2] = \begin{bmatrix} w^\star\cdot\frac{\cos(\alpha)}{\sqrt{\lambda_1}}, & w^\star\cdot\frac{\sin(\alpha)}{\sqrt{\lambda_2}} \\ w'\cdot\frac{\sin(\alpha)}{\sqrt{\lambda_1}}, & w'\frac{-\cos(\alpha)}{\sqrt{\lambda_2}}\end{bmatrix}\cdot\begin{bmatrix}\cos\theta & , \sin(\theta) \\ \sin\theta & , -\cos(\theta)\end{bmatrix} \tag{95}$$

To summarize, using arguments in Lemma J.3 and the fact that $w^\star = \mathbf{1}_{d_{\text{in}}}/\sqrt{d_{\text{in}}}$, we can afford to focus on just two rank two matrices $\Sigma_{A_\mathcal{W}}, \widetilde{\Sigma}_\mathcal{W}$ in the operation: $\text{SVD}_U(\Sigma_A^{-1/2})\widetilde{\Sigma}\Sigma_A^{-1/2}$. The other singular vectors from the SVD only impact directions that span $\mathcal{W}_\perp$, and the singular vectors obtained by considering only the rank 2 matrices lie only in the space of $\mathcal{W}$.

Part-II:

From the previous part we obtained forms of ϕ_1, ϕ_2 in terms of: $\lambda_1, \lambda_2, \alpha, \theta$, all of which are fully specified by the SVD of $\Sigma_{A_\mathcal{W}}$ and $\widetilde{\Sigma}_\mathcal{W}$. If we define $\tau := \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_2}}$, we can evaluate c_1, c_2, c_3, c_4 as:

$$c_1 = \frac{\cot(\alpha)}{\tau} + \tan(\theta) \tag{96}$$

$$c_2 = -1 + \frac{\cot(\alpha)\tan(\theta)}{\tau} \tag{97}$$

$$c_3 = \frac{1}{\tau} - \cot(\alpha)\tan(\theta) \tag{98}$$

$$c_4 = \frac{\tan(\theta)}{\tau} + \cot(\alpha) \tag{99}$$

Since $0 \leqslant \alpha, \theta \leqslant \pi/2$, and $\tau > 0$, we conclude that $c_1, c_4 \geqslant 0$. In Lemma J.8, we show that when $\frac{\gamma}{\sqrt{d_{\text{sp}}}} \ll 1$ and $\sigma_{\text{sp}_1} \leqslant \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$, $c_2 < 0$, and $c_3 > 0$. Now, we are ready to begin proofs for our claims on the amplification factors, *i.e.* on the ratios c_1/c_3, $|c_2/c_4|$. We will first show conditions on $|c1/c3|$, followed by those on $|c_2/c_4|$. For each of these conditions we will rely on the forms for c_1, c_2, c_3, c_4 derived in the previous part, in terms of α, θ, τ (where $0 \leqslant \alpha, \theta \leqslant \pi/2$). We will also rely on some lemmas that characterize the behavior of α, θ and τ as we vary σ_{sp} and keep all other problem parameters fixed. We defer the full proof of these lemmas to later sections. For this part of the proof, we also define additional notations: $\alpha_0, \tau_0, \theta_0$ as the quantities α, τ, θ respectively, when $\sigma_{\text{sp}} = 0$.

Lower bound on $c1/c3$.

From Lemma J.6, Lemma J.5, we know that for $\sigma_{\text{sp}_1} \leqslant \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$, the following conditions are satisfied:

- $\cot\alpha\tan(\theta) \leqslant \frac{9\gamma\cot(\alpha_0)}{\sqrt{d_{\text{sp}}}\tau_0}$. Furthermore, from Lemma J.6, we can also verify that ${}^{9\gamma\cot(\alpha_0)}/\sqrt{d_{\text{sp}}} \ll 1$ (as $\cot\alpha_0 \simeq {}^{\gamma}/\sqrt{d_{\text{sp}}}$).
- We know $\tau \leqslant p\tau_0$ where $p \in (1, 2)$ (from Lemma J.5). Thus, $\frac{1}{\tau} - \cot\alpha\tan(\theta) \geqslant \frac{1}{p\tau_0} - \frac{9\gamma\cot(\alpha_0)}{\sqrt{d_{\text{sp}}}\tau_0} > 0$, since $\frac{\gamma^2}{d_{\text{sp}}} \ll 1$ and $p < 2$. Thus, c_1/c_3 remains positive under our conditions on problem parameters.
- $\exists\sigma_{\text{sp}} \geqslant \sigma_{\text{sp}_1}$, such that $\tan(\theta) \geqslant \frac{5\gamma}{\sqrt{d_{\text{sp}}}\tau_0}$.

Now, since $\tau \geqslant \tau_0$ (see Lemma J.7), we can conclude that:

$$c_1/c_3 = \frac{{}^{\cot(\alpha)}/_{\tau} + \tan(\theta)}{{}^{1}/_{\tau} - \cot(\alpha)\tan(\theta)} \tag{100}$$

$$\geqslant \frac{\tan(\theta)}{{}^{1}/_{\tau_0} - \cot(\alpha)\tan(\theta)} \tag{101}$$

$$\geqslant \frac{\tan(\theta)}{{}^{1}/_{\tau_0} - \frac{9\gamma\cot(\alpha_0)}{\sqrt{d_{\text{sp}}}\tau_0}} \tag{102}$$

$$\geqslant \frac{\tan(\theta)}{{}^{1}/_{\tau_0}} \geqslant \frac{{}^{5\gamma}/\sqrt{d_{\text{sp}}}\tau_0}{{}^{1}/_{\tau_0}} = \frac{5\gamma}{\sqrt{d_{\text{sp}}}} \tag{103}$$

Thus some amplification on ϕ_1 is guaranteed as long as there is sufficient noise on the distribution of the spurious feature x_{sp} in the target domain.

Upper bound on $c1/c3$.

Now, we show that for the amplification on ϕ_1 is bounded when the noise on x_{sp} is not too high in target. From the same conditions on σ_{sp} as in the previous part, we know that $|c_1/c_3|$:

$$c_1/c_3 = \frac{\cot(\alpha)/\tau + \tan(\theta)}{1/\tau - \cot(\alpha)\tan(\theta)} \tag{104}$$

$$\leqslant \frac{\cot(\alpha_0)/\tau_0 + \tan(\theta)}{1/\tau - \cot(\alpha_0)\tan(\theta)} \tag{105}$$

$$\leqslant \frac{\cot(\alpha_0)/\tau_0 + 9\gamma/\sqrt{d_{\text{sp}}}\tau_0}{1/\tau - 9\gamma\cot(\alpha_0)/\sqrt{d_{\text{sp}}}\tau_0} \tag{106}$$

$$\leqslant \frac{\cot(\alpha_0)/\tau_0 + 9\gamma/\sqrt{d_{\text{sp}}}\tau_0}{1/p\tau_0 - 9\gamma\cot(\alpha_0)/\sqrt{d_{\text{sp}}}\tau_0} \tag{107}$$

$$= \frac{\cot(\alpha_0) + 9\gamma/\sqrt{d_{\text{sp}}}}{1/p - 9\gamma\cot(\alpha_0)/\sqrt{d_{\text{sp}}}} \tag{108}$$

where the first inequality uses $\tau_0 \leqslant \tau$ (see Lemma J.7) and $\cot(\alpha_0) \geqslant \cot(\alpha),\ \forall\sigma_{\text{sp}} > 0$ (see Lemma J.6), the second inequality uses Lemma J.5 which upper bounds $\tan(\theta)$ with $9\gamma/\sqrt{d_{\text{sp}}}\tau_0$, and the last inequality uses $\tau \leqslant p\tau_0$ for some $2 > p > 1$ (see Lemma J.5). Note that the final equality is a positive constant which is $\Theta(10\gamma p/\sqrt{d_{\text{sp}}})$ since $\cot(\alpha_0) = \Theta(\gamma/\sqrt{d_{\text{sp}}})$ (see Lemma J.6) when $\gamma/\sqrt{d_{\text{sp}}} \ll 1$. Since $p < 2$ from Lemma J.5, we can conclude that $c_1/c_3 \leqslant 20\gamma/\sqrt{d_{\text{sp}}}$. This upper bounds the amplification on ϕ_1 in Theorem H.12.

The lower and upper bounds on c_1/c_3 predominantly depend on the bounded nature of the noise on x_{sp} in target, *i.e.* when σ_{sp} is bounded, it implies that $\tan(\theta)$ and τ cannot be too large as compared to their values at no noise ($\sigma_{\text{sp}} = 0$). Next, we will verify the amplification claims on $|c_2/c_4|$.

Lower bound on $|c2/c4|$.

$$|c_2/c_4| = \frac{|-1 + \cot(\alpha)\tan(\theta)/\tau|}{|\tan(\theta)/\tau + \cot(\alpha)|} \tag{109}$$

$$\geqslant \frac{1 - \cot(\alpha_0)\tan(\theta)/\tau}{\tan(\theta)/\tau + \cot(\alpha_0)} \tag{110}$$

$$\geqslant \frac{1 - \cot(\alpha_0)9\gamma/\sqrt{d_{\text{sp}}}\tau_0\tau}{9\gamma/\sqrt{d_{\text{sp}}}\tau_0\tau + \cot(\alpha_0)} \tag{111}$$

$$\geqslant \frac{1 - \cot(\alpha_0)9\gamma/\sqrt{d_{\text{sp}}}\tau_0^2}{9\gamma/\sqrt{d_{\text{sp}}}\tau_0^2 + \cot(\alpha_0)} \tag{112}$$

$$= \frac{\tan(\alpha_0) - 9\gamma/\tau_0^2\sqrt{d_{\text{sp}}}}{9\gamma/\tau_0^2\sqrt{d_{\text{sp}}} + 1} \tag{113}$$

where the first inequality uses the condition $\cot(\alpha_0) \geqslant \cot(\alpha)$ (see Lemma J.6), and the second inequality uses $\tan(\theta) \leqslant 9\gamma/\sqrt{d_{\text{sp}}}\tau_0$ (see Lemma J.5). The final inequality use the condition $\tau \geqslant \tau_0$ (see Lemma J.7).

Let us now parse the final expression in the lower bound on $|c_2/c_4|$. When $\gamma/\sqrt{d_{\text{sp}}} \ll 1$, for *e.g.*, $\gamma/\sqrt{d_{\text{sp}}} \leqslant 0.2$ (as satisfied by our problem parameters in Example 1), then we can show that $\sqrt{d_{\text{sp}}}/\gamma(1+\epsilon_0) \geqslant 1/\cot(\alpha_0) \geqslant \sqrt{d_{\text{sp}}}/\gamma(1-\epsilon_0)$ for some small $\epsilon_0 > 0$ (see Lemma J.6). Further, we also know that $(1+\epsilon_1)\sqrt{d_{\text{sp}}}/\gamma \geqslant \tau_0(1-\epsilon_1)\sqrt{d_{\text{sp}}}/\gamma$ for some small $\epsilon_1 > 0$ (see Lemma J.7). Substituting these conditions in the final equality, we get:

$$|c_2/c_4| \geqslant \frac{(1-\epsilon_0)\sqrt{d_{\text{sp}}}/\gamma - \gamma^2(1-\epsilon_1)^2/d_{\text{sp}} \cdot 9\gamma/\sqrt{d_{\text{sp}}}}{\gamma^2(1-\epsilon_1)^2/d_{\text{sp}} \cdot 9\gamma/\sqrt{d_{\text{sp}}} + 1} \geqslant (1-\epsilon)\frac{\sqrt{d_{\text{sp}}}}{\gamma} \tag{114}$$

when $\gamma \ll \epsilon$, where $\epsilon > 0$ is some small positive constant.

Thus, with bounds on $|c_2/c_4|$, and c_1/c_3 (in Part-II) that amplify the effective margin in ϕ_1, ϕ_2, along with the claim on $\phi_j, \forall j \geqslant 3$ lying in the null space of $w_{\text{in}}, w_{\text{sp}}$ we have completed all parts of the claims in Theorem H.12. $\square$

H.6.4. PROOF OF COROLLARY H.4

Corollary H.13 (CL improves OOD error over ERM but is still imperfect)**.** *Under the conditions of Theorem H.12 the target accuracy of CL is at least* $0.5 \cdot \text{erfc}\left(-{c_1}/{(\omega c_3)} \cdot {\gamma}/{(\sqrt{2}\cdot\sigma_{\text{sp}})}\right)$*, and at most* $0.5 \cdot \text{erfc}\left(-{c_1}/{c_3} \cdot {\gamma}/{(\sqrt{2}\cdot\sigma_{\text{sp}})}\right)$*. Note that since* $c_1/c_3 \geqslant {5\gamma}/{\sqrt{d_{\text{sp}}}}$*, the lower bound is strictly better than ERM when* $1 \leqslant \omega \leqslant 5$ *which holds when* σ_2 *is small enough.*

Proof. Recall from Theorem H.12, all ϕ_j, for $j \geqslant 3$, lie in the null space of w_{in} and w_{sp}. Since, the predictive features are strictly contained in the rank t space spanned by w_{in} and w_{sp}, without loss of generality we can restrict ourselves to the case where $k = 2$, and when doing training a head $h = [h_1, h_2]^\top \in \mathbb{R}^2$ over contrastive pretrained representations using source labeled data, we get the following max margin solution:

$$h_1 = c_1 \cdot \gamma + c_3 \cdot \sqrt{d_{\text{sp}}} \tag{115}$$
$$h_2 = c_2 \cdot \gamma + c_4 \cdot \sqrt{d_{\text{sp}}} \tag{116}$$

Without loss of generality we can divide both h_1 and h_2 by h_1 and get the final classifier to be $\phi_1 + \frac{h_2}{h_1} \cdot \phi_2$:

$$(c_1 w_{\text{in}} + c_3 w_{\text{sp}}) + \frac{h_2}{h_1} \cdot (c_2 w_{\text{in}} + c_4 w_{\text{sp}}) \tag{117}$$
$$= (c_1 w_{\text{in}} + c_3 w_{\text{sp}}) + \frac{(c_2\gamma + c_4\sqrt{d_{\text{sp}}})}{(c_1\gamma + c_3\sqrt{d_{\text{sp}}})} \cdot (c_2 w_{\text{in}} + c_4 w_{\text{sp}}) \tag{118}$$

From the final part of Theorem H.12, we argued that: $(1-\epsilon)\sqrt{d_{\text{sp}}}/\gamma \leqslant |c_2/c_4| \leqslant \sqrt{d_{\text{sp}}}/\gamma$, where ϵ is a small positive constant $0 \leqslant \epsilon \leqslant 1$. Note that, c_2 is negative and c_1, c_3, c_4 are positive. Hence,

$$-(1-\epsilon)\frac{\sqrt{d_{\text{sp}}}}{\gamma} \geqslant \frac{c_2}{c_4} \geqslant -\frac{\sqrt{d_{\text{sp}}}}{\gamma}$$

$$\implies 0 \leqslant \frac{(c_2\gamma + c_4\sqrt{d_{\text{sp}}})}{(c_1\gamma + c_3\sqrt{d_{\text{sp}}})} \leqslant \frac{\epsilon c_4\sqrt{d_{\text{sp}}}}{c_1\gamma + c_3\sqrt{d_{\text{sp}}}} \tag{119}$$

Now, from Lemma J.9, we can derive the target accuracy of the classifier h on top of CL representations to be the following where $\beta = {(c_2\gamma + c_4\sqrt{d_{\text{sp}}})}/{(c_1\gamma + c_3\sqrt{d_{\text{sp}}})}$:

$$0.5\,\text{erfc}\left(-\frac{c_1 + \beta c_2}{c_3 + \beta c_4} \cdot \frac{\gamma}{\sqrt{2}\sigma_{\text{sp}}}\right) \tag{120}$$

Upper bound on target accuracy:

Note that $\beta = 0$, when $\epsilon = 0$. Hence, the best accuracy that we can get is $0.5\,\text{erfc}\left(-{c_1}/{c_3} \cdot {\gamma}/{\sqrt{2}\sigma_{\text{sp}}}\right)$. From Theorem H.12, we know that $c_1/c_3 \geqslant {5\gamma}/{\sqrt{d_{\text{sp}}}}$. Thus, the upper bound of the target performance is at least $0.5\,\text{erfc}\left(-{5\gamma^2}/{\sqrt{2d_{\text{sp}}}\sigma_{\text{sp}}}\right)$ which is better than the performance of ERM classifier (see Theorem H.6). But, also note that the upper bound on $c_1/c_3 \leqslant {20\gamma}/{\sqrt{d_{\text{sp}}}}$, which tells us that while c_1/c_3 scales up the effective margin, it does not solve the problem fully, *i.e*. the target accuracy is still not 100%. We will revisit this argument in the proof of STOC.

Lower bound on target accuracy:

As β increases the accuracy worsens. But, we have an upper bound on β that scales with ϵ. Now, for all sufficiently small $\epsilon \geqslant 0$, $\exists \omega > 1$ such that:

$$\frac{c_1 + \beta c_2}{c_3 + \beta c_4} \geqslant \frac{1}{\omega}\frac{c_1}{c_3} \tag{121}$$

When $\omega \leqslant \sqrt{d_{\text{sp}}}/\gamma$, the lower bound on target accuracy will become strictly better than ERM. Under our problem parameters in Example 1, $\omega = 4$ satisfies the constraint above. $\square$

H.7. Analysis of STOC: Formal Statement of Theorem H.5

Recall ERM solution over contrastive pretraining. We showed that without loss of generality when k (the output dimensionality of Φ) is greater than 2, we can restrict k to 2 and the Φ can be denoted as $[\phi_1, \phi_2]^\top$ where $\phi_1 = c_1 w^\star + c_3 w_{\text{sp}}$ and $\phi_2 = c_2 w^\star + c_4 w_{\text{sp}}$. The ERM solution of the linear head is then given by $h_1, h_2 \in \mathbb{R}$:

$$h_1 = c_1 \cdot \gamma + c_3 \cdot \sqrt{d_{\text{sp}}}\,, \quad \text{and} \quad h_2 = c_2 \cdot \gamma + c_4 \cdot \sqrt{d_{\text{sp}}}\,. \tag{122}$$

STOC performs self-training of the linear head over the CL solution. Before introducing the result, we need some additional notation. Let h^t denote the solution of the linear head at iterate t. Without loss of generality, assume that the coefficients in $\phi_1 = c_1 w_{\text{in}} + c_3 w_{\text{sp}}$ and $\phi_2 = c_2 w_{\text{in}} + c_4 w_{\text{sp}}$ are such that c_2 is positive and c_1, c_3, and c_4 are negative. Moreover, for simplicity of exposition, assume that $|c_4| > |c_3|$.

Theorem H.14. *Under the conditions of Corollary H.13, when* $\left|\frac{c_2}{c_4}\right| \geqslant 2 \cdot \frac{\sigma_{\text{sp}}}{\gamma} \cdot p(|c_4 \sigma_{\text{sp}}|, 0.5\,|c_1\gamma|) + \frac{c_1}{c_4}$ *(for p defined in (208)), the target accuracy of ST over CL is lower bounded by* $0.5 \cdot \text{erfc}\left(-\left|{c_2}/{c_4}\right| \cdot {\gamma}/{(\sqrt{2}\sigma_2)}\right) \geqslant 0.5 \cdot \text{erfc}\left(-{\sqrt{d_2}}/{(\sqrt{2}\sigma_2)}\right)$.

Proof. First, we create an outline of the proof. We argue about the updates of h^t showing that both h_1^t and h_2^t increase with $|h_2^t|$ becoming greater than $|h_1^t|$ for some large t. Then we show that $|h_2^t| \geqslant |h_1^t|$ is sufficient to obtain near-perfect target generalization.

Part 1. Recall the loss of used for self-training of h:

$$\begin{aligned}
\mathcal{L}_{\text{st}}(h) &= \mathbb{E}_{\mathrm{P_T}(x)}\left[\ell(h^\top \Phi x, \text{sgn}(h^\top \Phi x))\right] &(123)\\
&= \mathbb{E}_{\mathrm{P_T}(x)}\left[\exp\left(-\left|h^\top \Phi x\right|\right)\right] &(124)\\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[\exp\left(-\left|c_1\gamma h_1 + c_2\gamma h_2 + (c_3\sigma_{\text{sp}} h_1 + c_4 \sigma_{\text{sp}} h_2)\cdot z\right|\right)\right]\,. &(125)
\end{aligned}$$

Define $\mu_t = c_1\gamma h_1^t + c_2\gamma h_2^t$ and $\sigma_t = c_3\sigma_{\text{sp}} h_1^t + c_4\sigma_{\text{sp}} h_2^t$. With this notation, we can re-write the loss in (125) as $\mathcal{L}_{\text{st}}(h^t) = \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_t^2)}\left[\exp\left(-\left|\mu_t + z\right|\right)\right]$.

Similar to the the treatment in Theorem H.7, we now derive a closed-form expression of $\mathcal{L}_{\text{st}}(h^t)$ in Lemma J.10:

$$\mathcal{L}_{\text{st}}(h^t) = \frac{1}{2}\left(\exp\left(\frac{\sigma_t^2}{2} - \mu_t\right)\cdot \text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right) + \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right)\cdot\text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\right)\,. \tag{126}$$

Define:

$$\begin{aligned}
A_1(\mu_t, \sigma_t) &= \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right)\cdot\text{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\,, &(127)\\
A_2(\mu_t, \sigma_t) &= \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right)\cdot\text{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\,, &(128)\\
A_3(\mu_t, \sigma_t) &= \frac{2\sqrt{2}}{\sqrt{\pi}}\exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right)\,. &(129)
\end{aligned}$$

Let $\widetilde{h}^{t+1}$ denote the un-normalized gradient descent update at iterate $t+1$. We have:

$$\widetilde{h}^{t+1} = h^t - \eta \cdot \frac{\partial \mathcal{L}_{\text{st}}(h^t)}{\partial h}\,. \tag{130}$$

Now we will individually argue about the update of $\widetilde{h}^{t+1}$. First, we have:

$$\widetilde{h}_1^{t+1} = h_1^t - \eta \cdot \frac{\partial \mathcal{L}_{\mathrm{st}}(h^t)}{\partial h_1}$$
$$\widetilde{h}_1^{t+1} = h_1^t - \eta \cdot \underbrace{\left[A_1 \cdot (\sigma_t c_3 \sigma_{\mathrm{sp}} - c_1\gamma) + A_2 \cdot (\sigma_t c_3 \sigma_{\mathrm{sp}} + c_1\gamma) - A_3 c_3 \sigma_{\mathrm{sp}}\right]}_{\delta_1} . \tag{131}$$

and second, we have:

$$\widetilde{h}_2^{t+1} = h_2^t - \eta \cdot \frac{\partial \mathcal{L}_{\mathrm{st}}(h^t)}{\partial h_2}$$
$$\widetilde{h}_2^{t+1} = h_2^t - \eta \cdot \underbrace{\left[A_1 \cdot (\sigma_t c_4 \sigma_{\mathrm{sp}} - c_2\gamma) + A_2 \cdot (\sigma_t c_4 \sigma_{\mathrm{sp}} + c_2\gamma) - A_3 c_4 \sigma_{\mathrm{sp}}\right]}_{\delta_2} . \tag{132}$$

We will now argue the conditions under which h_2^{t+1} increases till its value reaches $1/\sqrt{2}$. In particular, we will argue that when h_2^t is negative, the norm $|h_2^t|$ decreases and when h_2^t becomes positive, then its norm increases. We show that the following three conditions are sufficient to argue the increasing value of h_2^t: for all t, we have (i) $\mu_t \geqslant \mu_c$ and $|\sigma_t| < \sigma_c$ for constant $\mu_c = |c_1 \cdot \gamma|/2$ and $\sigma_c = |c_4 \sigma_{\mathrm{sp}}|$; (ii) $\delta_2 < 0$; (iii) $|\delta_2| \geqslant \delta_1$. In Lemma H.16, we argue that our assumption on the initialization of the backbone learned with BT implies the previous three conditions.

Case-1. When h_2^t is negative (and after the update, it remains negative). Then we want to argue the following:

$$\frac{(h_2^t - \eta\delta_2)^2}{(h_2^t - \eta\delta_2)^2 + (h_1^t - \eta\delta_1)^2} \leqslant (h_2^t)^2 \tag{133}$$
$$\Rightarrow \quad \frac{(h_2^t - \eta\delta_2)^2}{(h_2^t)^2} \leqslant (h_2^t - \eta\delta_2)^2 + (h_1^t - \eta\delta_1)^2 \tag{134}$$
$$\Rightarrow \quad \frac{{h_2^t}^2 + \eta^2\delta_2^2 - 2\eta\delta_2 h_2^t}{(h_2^t)^2} \leqslant {h_2^t}^2 + \eta^2\delta_2^2 - 2\eta h_2^t\delta_2 + {h_1^t}^2 + \eta^2\delta_1^2 - 2\eta h_1^t\delta_1 \tag{135}$$
$$\Rightarrow \quad 1 + \frac{\eta^2\delta_2^2 - 2\eta\delta_2 h_2^t}{(h_2^t)^2} \leqslant 1 + \eta^2\delta_2^2 - 2\eta h_2^t\delta_2 + \eta^2\delta_1^2 - 2\eta h_1^t\delta_1 \tag{136}$$
$$\Rightarrow \quad \eta^2\delta_2^2 - 2\eta\delta_2 h_2^t \leqslant \left[\eta^2\delta_2^2 - 2\eta h_2^t\delta_2 + \eta^2\delta_1^2 - 2\eta h_1^t\delta_1\right](h_2^t)^2 \tag{137}$$
$$\Rightarrow \quad \eta^2\delta_2^2(h_1^t)^2 - 2\eta\delta_2 h_2^t(h_1^t)^2 \leqslant \eta^2\delta_1^2(h_2^t)^2 - 2\eta h_1^t\delta_1(h_2^t)^2 \tag{138}$$
$$\Rightarrow \quad \eta^2\delta_2^2(h_1^t)^2 - \eta^2\delta_1^2(h_2^t)^2 \leqslant 2\eta\delta_2 h_2^t(h_1^t)^2 - 2\eta h_1^t\delta_1(h_2^t)^2 \tag{139}$$
$$\Rightarrow \quad \left[\eta\delta_2(h_1^t) - \eta\delta_1(h_2^t)\right]\left[\eta\delta_2(h_1^t) + \eta\delta_1(h_2^t)\right] \leqslant 2h_2^t h_1^t\left[\eta\delta_2(h_1^t) - \eta\delta_1(h_2^t)\right] \tag{140}$$
$$\Rightarrow \quad \left[\eta\delta_2(h_1^t) + \eta\delta_1(h_2^t)\right] \leqslant 2h_2^t h_1^t \tag{141}$$

Since $\delta_2 < 0$, $|\delta_2| \geqslant |\delta_1|$ and $h_2^t < h_1^t < 0$, we have $[\eta\delta_2(h_1^t) - \eta\delta_1(h_2^t)]$ as positive. This implies inequality (140) to (141) and for small enough η, (141) will continue to hold true.

Case-2. When h_2^t is positive but less than $1/\sqrt{2}$. Then we want to argue the following:

$$
\begin{aligned}
& & \frac{(h_2^t - \eta\delta_2)^2}{(h_2^t - \eta\delta_2)^2 + (h_1^t - \eta\delta_1)^2} &\geqslant (h_2^t)^2 & (142)\\
&\Rightarrow & \frac{(h_2^t - \eta\delta_2)^2}{(h_2^t)^2} &\geqslant (h_2^t - \eta\delta_2)^2 + (h_1^t - \eta\delta_1)^2 & (143)\\
&\Rightarrow & \frac{{h_2^t}^2 + \eta^2\delta_2^2 - 2\eta\delta_2 h_2^t}{(h_2^t)^2} &\geqslant {h_2^t}^2 + \eta^2\delta_2^2 - 2\eta h_2^t\delta_2 + {h_1^t}^2 + \eta^2\delta_1^2 - 2\eta h_1^t\delta_1 & (144)\\
&\Rightarrow & 1 + \frac{\eta^2\delta_2^2 - 2\eta\delta_2 h_2^t}{(h_2^t)^2} &\geqslant 1 + \eta^2\delta_2^2 - 2\eta h_2^t\delta_2 + \eta^2\delta_1^2 - 2\eta h_1^t\delta_1 & (145)\\
&\Rightarrow & \eta^2\delta_2^2 - 2\eta\delta_2 h_2^t &\geqslant \left[\eta^2\delta_2^2 - 2\eta h_2^t\delta_2 + \eta^2\delta_1^2 - 2\eta h_1^t\delta_1\right](h_2^t)^2 & (146)\\
&\Rightarrow & \eta^2\delta_2^2(h_1^t)^2 - 2\eta\delta_2 h_2^t(h_1^t)^2 &\geqslant \eta^2\delta_1^2(h_2^t)^2 - 2\eta h_1^t\delta_1(h_2^t)^2 & (147)\\
&\Rightarrow & \eta^2\delta_2^2(h_1^t)^2 - \eta^2\delta_1^2(h_2^t)^2 &\geqslant 2\eta\delta_2 h_2^t(h_1^t)^2 - 2\eta h_1^t\delta_1(h_2^t)^2 & (148)\\
&\Rightarrow & \left[\eta\delta_2(h_1^t) - \eta\delta_1(h_2^t)\right]\left[\eta\delta_2(h_1^t) + \eta\delta_1(h_2^t)\right] &\geqslant 2h_2^t h_1^t\left[\eta\delta_2(h_1^t) - \eta\delta_1(h_2^t)\right] & (149)\\
&\Rightarrow & \left[\eta\delta_2(h_1^t) + \eta\delta_1(h_2^t)\right] &\geqslant 2h_2^t h_1^t & (150)
\end{aligned}
$$

Since $\delta_2 < 0$, $|\delta_2| \geqslant |\delta_1|$, $h_1^t \leqslant -1/\sqrt{2}$ and $0 < h_2^t < 1/\sqrt{2}$, we have $[\eta\delta_2(h_1^t) - \eta\delta_1(h_2^t)]$ as positive. This implies inequality (149) to (150). Focusing on (150), we note that $h_1^t \cdot \delta_2$ is positive and greater in magnitude than $h_2^t \cdot \delta_1$. Moreover, since $h_2^t h_1^t$ is negative, (150) will continue to hold true.

Now, when h_2^t is positive and greater than $1/\sqrt{2}$, then h_2^t will stay in that region. Convergence of STOC together with conditions of convergence as in Lemma H.15 will imply that the at convergence h_2^t will remain greater than $1/\sqrt{2}$, such that $\frac{h_1^{t_c}}{h_2^{t_c}} = \frac{\delta_1}{\delta_2}$. Now we bound the target error of STOC.

Part 2. To bound the accuracy at any iterate t when $h_2^t \geqslant 1/\sqrt{2}$, we have from Lemma J.9:

$$
\mathbb{E}_{\mathrm{P_T}}\left[y \cdot \left({h^t}^\top \phi_{\mathrm{cl}} x\right) > 0\right] = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[z > -\frac{c_1\gamma h_1^t + c_2\gamma h_2^t}{|c_3\sigma_{\mathrm{sp}} h_1^t + c_4\sigma_{\mathrm{sp}} h_2^t|}\right]. \tag{151}
$$

We now upper bound and lower bound the fraction $\frac{c_1\gamma h_1^t + c_2\gamma h_2^t}{|c_3\sigma_{\mathrm{sp}} h_1^t + c_4\sigma_{\mathrm{sp}} h_2^t|}$ in RHS in (151): (i) $c_1\gamma h_1^t + c_2\gamma h_2^t \geqslant c_2\gamma h_2^t$ since both $c_1\gamma h_1^t$ and $c_2\gamma h_2^t$ have same sign; (ii) $|c_3\sigma_{\mathrm{sp}} h_1^t + c_4\sigma_{\mathrm{sp}} h_2^t| \leqslant |c_4\sigma_{\mathrm{sp}} h_2^t|$ because $|c_4\sigma_{\mathrm{sp}} h_2^t| \geqslant |c_3\sigma_{\mathrm{sp}} h_1^t|$ and they have opposite signs. Hence, from (151), we have:

$$
\mathbb{E}_{\mathrm{P_T}}\left[y \cdot \left({h^t}^\top \phi_{\mathrm{cl}} x\right) > 0\right] = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[z > -\frac{c_2\gamma h_2^t}{|c_4\sigma_{\mathrm{sp}} h_2^t|}\right] = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[z > -\frac{c_2\gamma}{|c_4\sigma_{\mathrm{sp}}|}\right]. \tag{152}
$$

Substituting the definition of erfc, the expression (152) gives us the required lower bound on the target accuracy.

□

Lemma H.15 (Convergence of STOC)**.** *Assume the gradient updates as in* (131) *and* (132)*. Then STOC converges at* $t = t_c$ *when* $\frac{h_1^{t_c}}{h_2^{t_c}} = \frac{\delta_1}{\delta_2}$. *For* $t > t_c$, (131) *and* (132) *make no updates to the linear* h.

Proof. When the gradient updates δ_1 and δ_2 are such that h_1^{t+1} matches h_1^t, we have convergence of STOC.

$$\begin{aligned}
& & \frac{(h_2^t-\eta\delta_2)^2}{(h_2^t-\eta\delta_2)^2+(h_1^t-\eta\delta_1)^2} &= (h_2^t)^2 && (153)\\
&\Rightarrow & \frac{(h_2^t-\eta\delta_2)^2}{(h_2^t)^2} &= (h_2^t-\eta\delta_2)^2+(h_1^t-\eta\delta_1)^2 && (154)\\
&\Rightarrow & \frac{{h_2^t}^2+\eta^2\delta_2^2-2\eta\delta_2 h_2^t}{(h_2^t)^2} &= {h_2^t}^2+\eta^2\delta_2^2-2\eta h_2^t\delta_2+{h_1^t}^2+\eta^2\delta_1^2-2\eta h_1^t\delta_1 && (155)\\
&\Rightarrow & 1+\frac{\eta^2\delta_2^2-2\eta\delta_2 h_2^t}{(h_2^t)^2} &= 1+\eta^2\delta_2^2-2\eta h_2^t\delta_2+\eta^2\delta_1^2-2\eta h_1^t\delta_1 && (156)\\
&\Rightarrow & \eta^2\delta_2^2-2\eta\delta_2 h_2^t &= \left[\eta^2\delta_2^2-2\eta h_2^t\delta_2+\eta^2\delta_1^2-2\eta h_1^t\delta_1\right](h_2^t)^2 && (157)\\
&\Rightarrow & \eta^2\delta_2^2(h_1^t)^2-2\eta\delta_2 h_2^t(h_1^t)^2 &= \eta^2\delta_1^2(h_2^t)^2-2\eta h_1^t\delta_1(h_2^t)^2 && (158)\\
&\Rightarrow & \eta^2\delta_2^2(h_1^t)^2-\eta^2\delta_1^2(h_2^t)^2 &= 2\eta\delta_2 h_2^t(h_1^t)^2-2\eta h_1^t\delta_1(h_2^t)^2 && (159)\\
&\Rightarrow & \left[\eta\delta_2(h_1^t)-\eta\delta_1(h_2^t)\right]\left[\eta\delta_2(h_1^t)+\eta\delta_1(h_2^t)\right] &= 2h_2^t h_1^t\left[\eta\delta_2(h_1^t)-\eta\delta_1(h_2^t)\right] && (160)
\end{aligned}$$

Thus either $[\eta\delta_2(h_1^t)-\eta\delta_1(h_2^t)] = 0$ or $[\eta\delta_2(h_1^t)+\eta\delta_1(h_2^t)] = 2h_2^t h_1^t$. Since η is such that $h_1-\eta\delta_1 < 0$, $[\eta\delta_2(h_1^t)+\eta\delta_1(h_2^t)] \neq 2h_2^t h_1^t$ implying that $[\eta\delta_2(h_1^t)-\eta\delta_1(h_2^t)] = 0$ giving us the required condition. □

Lemma H.16. *Under the initialization conditions assumed in Theorem H.14, for all t, we have: (i) $\mu_t \geqslant \mu_c$ and $|\sigma_t| \leqslant \sigma_c$ for constant $\mu_c = |c_1\cdot\gamma|/2$ and $\sigma_c = |c_4\sigma_{\text{sp}}|$; (ii) $\delta_2 < 0$; (iii) $|\delta_2| \geqslant \delta_1$, where $\delta_1 = A_1\cdot(\sigma_t c_3\sigma_{\text{sp}} - c_1\gamma) + A_2\cdot(\sigma_t c_3\sigma_{\text{sp}} + c_1\gamma) - A_3 c_3\sigma_{\text{sp}}$ and $\delta_2 = A_1\cdot(\sigma_t c_4\sigma_{\text{sp}} - c_2\gamma) + A_2\cdot(\sigma_t c_4\sigma_{\text{sp}} + c_2\gamma) - A_3 c_4\sigma_{\text{sp}}$ for A_1, A_2 and A_3 defined in (127), (128), and (129).*

Proof. Recall, $\mu_t = c_1\gamma h_1^t + c_2\gamma h_2^t$ and $\sigma_t = c_3\sigma_{\text{sp}}h_1^t + c_4\sigma_{\text{sp}}h_2^t$.

First, we argue that μ_t increases from the initialization value. Notice that $\mu_0 = c_1\gamma h_1^0 + c_2\gamma h_2^0$. Due to Corollary H.13, we have $h_2^0\cdot c_2 \ll c_1 h_1^0$ implying $\mu_0 \geqslant |c_1\gamma|/2$ as both c_1 and h_1^0 are of same sign and h_1^0 is close to -1. As h_2^t becomes positive since $c_2 >> c_1$, $c_2 h_2^t$ increases at a faster rate than the decrease in $c_1 h_1^t$ implying that $\mu_t \geqslant \mu_c$ continues to hold true. Since $|c_4| > |c_3|$, and both $|h_1^t|, |h_2^t| \leqslant 1$, we have $\sigma_t \leqslant |c_4\sigma_{\text{sp}}|$.

To argue (ii) and (iii), we use Lemma J.11 which provides an upper bound on $\frac{A_3 - A_1\sigma_t - A_2\sigma_t}{A_1 - A_2}$ as $p(\sigma_0, \mu_0)$ with p defined in (208). According to the expression of δ_2, we have:

$$\begin{aligned}
\delta_2 &= A_1\cdot(\sigma_t c_4\sigma_{\text{sp}} - c_2\gamma) + A_2\cdot(\sigma_t c_4\sigma_{\text{sp}} + c_2\gamma) - A_3 c_4\sigma_{\text{sp}} && (161)\\
&= (A_1\cdot\sigma_t + A_2\cdot\sigma_t - A_3)\, c_4\sigma_{\text{sp}} - c_2(A_1\gamma - A_2\gamma) && (162)\\
&= \left(\frac{(-A_1\cdot\sigma_t - A_2\cdot\sigma + A_3)}{(A_1 - A_2)} - \frac{c_2\gamma}{-c_4\sigma_{\text{sp}}}\right)(-c_4\sigma_{\text{sp}} * (A_1 - A_2)) && (163)\\
&\leqslant 0\,, && (164)
\end{aligned}$$

when $\frac{c_2\gamma}{(-c_4\sigma_{\text{sp}})} \geqslant p(|c_4\sigma_{\text{sp}}|, 0.5|c_1\gamma|)$. Similarly for (iii), putting in expressions for δ_1 and δ_2, we get: $\frac{c_2\gamma}{(-c_4\sigma_{\text{sp}})} \geqslant 2\cdot p(|c_4\sigma_{\text{sp}}|, 0.5|c_1\gamma|) + \frac{c_1\gamma}{c_4\sigma_{\text{sp}}}$.

□

H.8. Analysis for SSL

For SSL analysis, we argue that the projection learned by contrastive pretraining can significantly improve the generalization of the linear head learned on top, leaving little to no room for improvement for self-training. Our analysis leverages the margin-based bound for linear models from Kakade et al. (2008). Before introducing the result, we present some additional notation. Let $\text{Err}_D(w)$ denote 0-1 error of a classifier on a distribution D. Define 0-1 error with margin γ as $\widehat{\text{Err}}^{\gamma}(w) = \sum_{i=1}^{n} \frac{\mathbb{I}\left[y_i w^\top x_i \leqslant \gamma\right]}{n}$.

Theorem H.17 (Corollary 6 in Kakade et al. (2008))**.** *For all classifiers w and margin γ, we have with probability at least*

$1 - \delta$:

$$\mathrm{Err}_T(w) \leqslant \widehat{\mathrm{Err}}^{\gamma}(w) + 4\frac{B}{\gamma}\sqrt{\frac{1}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(\log_2(4B/\gamma))}{n}}\,, \tag{165}$$

where B is an upper bound on the ℓ_2 norm of the input points x.

When $\widetilde{\mathrm{Err}}^{\gamma}(w)$ is close to zero, the denominating term in RHS of (165) is $4\frac{B}{\gamma}\sqrt{\frac{1}{n}}$. SSL mainly reduces the B on the projected data by reducing the dependency from order $\sqrt{d}$ to $\sqrt{k}$ where k is the dimensionality of the output of ϕ. This reduction is the best possible in the setting where contrastive representations do not significantly lose the margin (separating classes) on the original input data, *i.e.*, γ does not drop too much. This is true in our theoretical analysis when the conditions in Theorem H.12 are satisfied. Intuitively, since the target data has only one predictive feature (along w_{in}), CL directly recovers this predictive feature since it is the predominant direction that minimizes invariance loss.

Moreover, in our setup, all the points are at the margin, and hence $\widetilde{\mathrm{Err}}^{\gamma}(w)$ will be zero or one. When training error is close to zero,

I. Limitations of Prior Work

I.1. Contrastive learning analysis

Prior works that analyze contrastive learning show that minimizers of the CL objective recover clusters in the augmentation graph, which weights pairs of augmentations with their probability of being sampled as a positive pair (HaoChen et al., 2021; Cabannes et al., 2023; Saunshi et al., 2022; Johnson et al., 2022). When there is no distribution shift in the downstream task, assumptions made on the graph in the form of consistency of augmentations with downstream labels, is sufficient to ensure that a linear probed head has good ID generalization. Under distribution shift, these assumptions are not sufficient and stronger ones are needed. *E.g.*, some works assume that same-domain/class examples are weighted higher that cross-class cross-domain pairs (HaoChen et al., 2022; Shen et al., 2022).

Using notation defined in (Shen et al., 2022), the assumption on the augmentation graph requires cross-class and same-domain weights (β) to be higher than cross-class and cross-domain weights (γ). It is unclear if examples from different classes in the same domain will be "connected" if strong spurious features exist in the source domain and augmentations fail to mask them completely (*e.g.*, image background may not be completely masked by augmentations but it maybe perfectly predictive of the label on source domain). In such cases, the linear predictor learnt over CL would fail to generalize OOD. In our toy setup as well, the connectivity assumption fails since on source x_{sp} is perfectly predictive of the label and the augmentations are imperfect, *i.e.*, augmentations do not mask x_{sp} and examples of different classes do not overlap in source (*i.e.*, $\beta = 0$). On the other hand, since x_{sp} is now random on target, augmentations of different classes may overlap, *i.e.*, $\gamma > 0$, thus breaking the connectivity assumption. This is also highlighted in our empirical findings of CL furnishing representations that do not fully enable linear transferability from source to target (see Sec. B). These empirical findings also call into question existing assumptions on data augmentations, highlighting that perfect linear transferability may not typically hold in practice. It is in this setting that we believe self-training can improve over contrastive learning by unlearning source-only features and improving linear transferability.

I.2. Self-training analysis

Some prior works on self-training view it as consistency regularization that constrain pseudolabels of original samples to be consistent with all their augmentations (Cai et al., 2021; Wei et al., 2020; Sohn et al., 2020). This framework abstracts the role played by the optimization algorithm and instead evaluates the global minimizer of a population objective that enforces consistency of pseudolabels. In addition, certain expansion assumptions on class-conditional distributions are needed to ensure that pseudolabels have good accuracy on source and target domains. This framework does not account for challenges involved in propagating labels iteratively. For *e.g.*, when augmentation distribution has long tails, the consistency of pseudolabels depends on the sampling frequency of "favorable" augmentations. As an illustration, consider our augmentation distribution in the toy setup in Sec. 3. If it were not uniform over dimensions, but instead something that was highly skewed, then a large number of augmentations need to be sampled for every data point to propagate pseudolabels successfully from source labeled samples to target unlabeled samples during self-training. This might hurt the performance of ST when we are optimizing for only finitely many iterations and over finitely many datapoints. This is why in our analysis we instead adopt the iterative analysis of self-training (Chen et al., 2020b).

J. Additional Lemmas

In this section we define some additional lemmas that we use in our theoretical analysis in H.

Lemma J.1 (Upper bound and lower bounds on erfc; Kschischang (2017))**.** *Define* $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \cdot \int_x^\infty \exp(-z^2) \cdot dz$. *Then we have:*

$$\frac{2}{\sqrt{\pi}} \cdot \frac{\exp(-x^2)}{x + \sqrt{x^2+2}} < \mathrm{erfc}(x) \leqslant \frac{2}{\sqrt{\pi}} \cdot \frac{\exp(-x^2)}{x + \sqrt{x^2 + 4/\pi}}$$

Lemma J.2 (invariance loss as product with operator L)**.** *The invariance loss for some* $\phi \in \mathbb{R}^d$ *is given as:* $2 \cdot \int_{\mathcal{A}} \phi(a) \cdot L(\phi)(a) \; \mathrm{dP}_{\mathsf{A}}$ *where the operator* L *is defined as:*

$$L(\phi)(a) = \phi(a) - \int_{\mathcal{A}} \frac{A_+(a, a')}{p_{\mathsf{A}}(a)} \cdot \phi(a') \; \mathrm{d}a'$$

Proof. The invariance loss for ϕ is given by:

$$\begin{aligned}
\mathbb{E}_{x\sim \mathrm{P_U}} \mathbb{E}_{a_1, a_2 \sim \mathrm{P_A}(\cdot|x)} (a_1^\top \phi - a_2^\top \phi)^2 &= 2\mathbb{E}_{x\sim \mathrm{P_U}} \mathbb{E}_{a\sim \mathrm{P_A}(\cdot|x)} \left[\phi(a)^2\right] \\
&\quad - 2\mathbb{E}_{a_1,a_2\sim A_+(\cdot,\cdot)} \left[\phi(a_1)\phi(a_2)\right] &(166)\\
= 2 \cdot \int_{\mathcal{A}} \phi(a)^2 \; \mathrm{dP}_{\mathsf{A}} - 2 \cdot \int_{\mathcal{A}} \phi(a) \left(\int_{\mathcal{A}} \frac{A_+(a, a_2)}{p_{\mathsf{A}}(a)} \cdot \phi(a_2) \; \mathrm{d}a_2 \right) \; \mathrm{dP}_{\mathsf{A}} & &(167)\\
= 2 \cdot \int_{\mathcal{A}} \phi(a) \cdot L(\phi)(a) \; \mathrm{dP}_{\mathsf{A}} & &(168)
\end{aligned}$$

□

Lemma J.3. *If* $\mathcal{W}$ *is the space spanned by* w_{in} *and* w_{sp}*, and* $\mathcal{W}_\perp$ *is the null space for* $\mathcal{W}$*, then for any* $u \in \mathcal{W}$ *and any* $v \in \mathcal{W}_\perp$*, the covariance along these directions* $\mathbb{E}_{a\sim \mathrm{P_A}}[a^\top u v^\top a] = 0$.

Proof: We can write the covariance over augmentations after we break down the augmentation a into two projections: $a = \Pi_{\mathcal{W}}(a) + \Pi_{\mathcal{W}_\perp}(a)$

$$\begin{aligned}
\mathbb{E}_{a\sim \mathrm{P_A}}[a^\top u v^\top a] &= \mathbb{E}_{a\sim \mathrm{P_A}} \left[\left(u^\top (\Pi_{\mathcal{W}}(a) + \Pi_{\mathcal{W}_\perp}(a))\right) \left(v^\top (\Pi_{\mathcal{W}}(a) + \Pi_{\mathcal{W}_\perp}(a))\right) \right] &(169)\\
&= \mathbb{E}_{a\sim \mathrm{P_A}} \left[\left(u^\top \Pi_{\mathcal{W}}(a)\right) \left(v^\top \Pi_{\mathcal{W}_\perp}(a)\right) \right] &(170)\\
&= u^\top \left(\mathbb{E}_{a\sim \mathrm{P_A}} \left[\Pi_{\mathcal{W}}(a) \Pi_{\mathcal{W}_\perp}(a)^\top \right] \right) v = 0 &(171)
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}_{a\sim \mathrm{P_A}} \left[\Pi_{\mathcal{W}}(a)\Pi_{\mathcal{W}_\perp}(a)^\top\right] = \mathbb{E}_{a\sim \mathrm{P_A}} \left[\Pi_{\mathcal{W}}(a)\right] \mathbb{E}_{a\sim \mathrm{P_A}} \left[\Pi_{\mathcal{W}_\perp}(a)\right]^\top$, since the noise in the null space of $\mathcal{W}$ is drawn independent of the component along $\mathcal{W}$, and furthermore the individual expectations evaluate to zero.

Lemma J.4. *For a* 2×2 *real symmetric matrix* $\begin{bmatrix} a, & b \\ c, & d \end{bmatrix}$ *the eigenvalues* λ_1, λ_2 *are given by the following expressions:*

$$\begin{aligned}
\lambda_1 &= (a + b + \delta)/2 \\
\lambda_2 &= (a + b - \delta)/2,
\end{aligned}$$

where $\delta = \sqrt{4c^2 + (a-b)^2}$. *Further, the eigenvectors are given by* $U = \begin{bmatrix} \cos(\theta), & \sin(\theta) \\ \sin(\theta), & -cos(\theta) \end{bmatrix}$, *where* $\tan(\theta)$ *is defined as follows:*

$$\tan(\theta) = \frac{b - a + \delta}{2c}$$

For full proof of the above statements see (Deledalle et al., 2017). Here, we will use these statements to arrive at closed form expressions for the eigenvalues and eigenvectors of Σ_A, $\widetilde{\Sigma}$ *and their approximations when* $\gamma \ll \sqrt{d_{\mathrm{sp}}}$, i.e. $\frac{\gamma}{\sqrt{d_{\mathrm{sp}}}} \leqslant \epsilon$, *where* ϵ *is a small positive constant (of the order of* ≈ 0.1 *for the problem parameters defined in Example 1).*

Proof. We can now substitute the above formulae with a, b, c, d taken from the expressions of Σ_A and $\widetilde{\Sigma}$, to get the following values: λ_1, λ_2 are the eigenvalues of Σ_A, with α determining the corresponding eigenvectors $[\cos(\alpha), \sin(\alpha)], [\sin(\alpha), -\cos(\alpha)]$; and $\widetilde{\lambda}_1, \widetilde{\lambda}_2$ are the eigenvalues of $\widetilde{\Sigma}$, with β determining the corresponding eigenvectors: $[\cos(\beta), \sin(\beta)], [\sin(\beta), -\cos(\beta)]$.

$$\begin{aligned}\lambda_1 = \frac{1}{8}\Bigg(&\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+\frac{\sigma_{\text{in}}^2}{3}\left(1-\frac{1}{d_{\text{in}}}\right)+\frac{d_{\text{sp}}}{2}+\frac{2\sigma_{\text{sp}}^2}{3}+\frac{1}{6}\\ &+\sqrt{\gamma^2 d_{\text{sp}}+\left(\left(\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+\frac{\sigma_{\text{in}}^2}{3}\left(1-\frac{1}{d_{\text{in}}}\right)\right)-\left(\frac{d_{\text{sp}}}{2}+\frac{2\sigma_{\text{sp}}^2}{3}+\frac{1}{6}\right)\right)^2}\Bigg)\end{aligned} \tag{172}$$

$$\begin{aligned}\lambda_2 = \frac{1}{8}\Bigg(&\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+\frac{\sigma_{\text{in}}^2}{3}\left(1-\frac{1}{d_{\text{in}}}\right)+\frac{d_{\text{sp}}}{2}+\frac{2\sigma_{\text{sp}}^2}{3}+\frac{1}{6}\\ &-\sqrt{\gamma^2 d_{\text{sp}}+\left(\left(\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+\frac{\sigma_{\text{in}}^2}{3}\left(1-\frac{1}{d_{\text{in}}}\right)\right)-\left(\frac{d_{\text{sp}}}{2}+\frac{2\sigma_{\text{sp}}^2}{3}+\frac{1}{6}\right)\right)^2}\Bigg)\end{aligned} \tag{173}$$

$$\widetilde{\lambda}_1 = \frac{1}{8}\left(\gamma^2+\frac{d_{\text{sp}}}{2}+\frac{\sigma_{\text{sp}}^2}{2}+\sqrt{\gamma^2 d_{\text{sp}}+\left(\gamma^2-\left(\frac{d_{\text{sp}}}{2}+\frac{\sigma_{\text{sp}}^2}{2}\right)\right)^2}\right) \tag{174}$$

$$\widetilde{\lambda}_2 = \frac{1}{8}\left(\gamma^2+\frac{d_{\text{sp}}}{2}+\frac{\sigma_{\text{sp}}^2}{2}-\sqrt{\gamma^2 d_{\text{sp}}+\left(\gamma^2-\left(\frac{d_{\text{sp}}}{2}+\frac{\sigma_{\text{sp}}^2}{2}\right)\right)^2}\right) \tag{175}$$

$$\begin{aligned}\tan(\alpha) = \frac{1}{\gamma\sqrt{d_{\text{sp}}}}\Bigg(&\frac{d_{\text{sp}}}{2}+\frac{2\sigma_{\text{sp}}^2}{3}+\frac{1}{6}-\left(\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+\frac{\sigma_{\text{in}}^2}{3}\left(1-\frac{1}{d_{\text{in}}}\right)\right)\\ &+\sqrt{\gamma^2 d_{\text{sp}}+\left(\left(\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+\frac{\sigma_{\text{in}}^2}{3}\left(1-\frac{1}{d_{\text{in}}}\right)\right)-\left(\frac{d_{\text{sp}}}{2}+\frac{2\sigma_{\text{sp}}^2}{3}+\frac{1}{6}\right)\right)^2}\Bigg)\end{aligned} \tag{176}$$

$$\tan(\beta) = \frac{1}{\gamma\sqrt{d_{\text{sp}}}}\left(\frac{d_{\text{sp}}}{2}+\frac{\sigma_{\text{sp}}^2}{2}-\gamma^2+\sqrt{\gamma^2 d_{\text{sp}}+\left(\gamma^2-\left(\frac{d_{\text{sp}}}{2}+\frac{\sigma_{\text{sp}}^2}{2}\right)\right)^2}\right) \tag{177}$$

For each of these quantities: $\lambda_1, \lambda_2, \tan(\alpha), \widetilde{\lambda}_1, \widetilde{\lambda}_2, \tan(\beta)$, we can directly apply the limit $\gamma/\sqrt{d_{\text{sp}}} \to 0$ to get the following expressions:

$$\lambda_1 \approx \frac{1}{8}\cdot\left(\gamma^2\left(1+\frac{1}{3d_{\text{in}}}\right)+d_{\text{sp}}+\frac{4}{3}\sigma_{\text{sp}}^2+\frac{1}{3}\right) \tag{178}$$

$$\tan(\alpha) \approx \frac{d_{\text{sp}}+\frac{4}{3}\cdot\sigma_{\text{sp}}^2-\gamma^2\cdot(1+1/3d_{\text{in}})+1/3}{\gamma\sqrt{d_{\text{sp}}}} \tag{179}$$

$$\widetilde{\lambda}_1 \approx \frac{1}{8}\cdot\left(\gamma^2+d_{\text{sp}}+\sigma_{\text{sp}}^2\right) \tag{180}$$

$$\tan(\beta) \approx \frac{\sigma_{\text{sp}}^2+d_{\text{sp}}-\gamma^2}{\gamma\sqrt{d_{\text{sp}}}} \tag{181}$$

□

Lemma J.5. *When $\gamma, \sigma_{\text{in}} \ll d_{\text{sp}}$ (conditions from Theorem H.12), we can show that $\exists \sigma_{\text{sp}_1}, \sigma_{\text{sp}_2}$ such that for the range of $\sigma_{\text{sp}_1} \leqslant \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$, $5\gamma/\tau_0\sqrt{d_{\text{sp}}} \leqslant \tan\theta \leqslant 9\gamma/\tau_0\sqrt{d_{\text{sp}}}$. Further, there exists $p \in (1, 2)$ such that $\tau \leqslant p\tau_0$. For the problem parameters defined in Example 1, ${\sigma_{\text{sp}}}_1^2 = 0.8$, and ${\sigma_{\text{sp}}}_2^2 = 1.5$ satisfies the conditions we need.*

Proof. Using (Stewart, 1993), we know that the singular vectors of a 2×2 asymmetric matrix $\begin{bmatrix} a, & b \\ c, & d \end{bmatrix}$, is $\begin{bmatrix} \cos\theta, & \sin\theta \\ \sin\theta, & -\cos\theta \end{bmatrix}$.

Here, $\tan(2\theta)$ is given by:

$$\tan(2\theta) = \frac{2ac + 2bd}{a^2 + b^2 - c^2 - d^2}$$

Now, substituting the values in the expression (94), we get:

$$\tan(2\theta) = \frac{2\tan(\alpha - \beta) \cdot (\widetilde{\lambda}_1 - \widetilde{\lambda}_2) \cdot \sqrt{\lambda_1\lambda_2}}{(\lambda_2\widetilde{\lambda}_1 - \lambda_1\widetilde{\lambda}_2) - (\lambda_1\widetilde{\lambda}_1 - \lambda_2\widetilde{\lambda}_2) \cdot \tan^2(\alpha - \beta)} \tag{182}$$

$$= \frac{2\tan(\alpha - \beta) \cdot (\widetilde{\lambda}_1/\widetilde{\lambda}_2 - 1) \cdot \sqrt{\lambda_1}/\sqrt{\lambda_2}}{(\widetilde{\lambda}_1/\widetilde{\lambda}_2 - \lambda_1/\lambda_2) - (\lambda_1\widetilde{\lambda}_1/\lambda_2\widetilde{\lambda}_2 - 1) \cdot \tan^2(\alpha - \beta)} \tag{183}$$

Controlling $\alpha - \beta$:

We will first note that α increases based on our arguments in Lemma J.6. Using similar arguments, we can also claim β increases, but the key point here is that due to the effect of augmentations α increases at a rate that is faster than rate of increase of β, specifically when σ_{sp} is not too large. We can see this by analyzing $\frac{\partial \tan(\alpha)}{\partial \sigma_{\text{sp}}}$ and comparing it with $\frac{\partial \tan(\beta)}{\partial \sigma_{\text{sp}}}$, in the region $\sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$ (essentially the region where we can approximate $\tan(\alpha), \tan(\beta)$ with their first order Taylor approximations). From the expressions for $\tan(\alpha), \tan(\beta)$ in Lemma J.4, under $\gamma, \sigma_{\text{in}} \ll \sqrt{d_{\text{sp}}}$, we get: $\frac{\partial \tan(\beta)}{\partial \sigma_{\text{sp}}} = \mathcal{O}(2\sigma_{\text{sp}}/\gamma\sqrt{d_{\text{sp}}})$, and $\frac{\partial \tan(\alpha)}{\partial \sigma_{\text{sp}}} = \mathcal{O}(8\sigma_{\text{sp}}/3\gamma\sqrt{d_{\text{sp}}})$. This establishes the fact that $\tan(\alpha-\beta)$ increases monotonically in some range for σ_{sp}, as long as $\gamma, \sigma_{\text{in}} \ll \sqrt{d_{\text{sp}}}$.

Controlling functions of $\lambda_1, \lambda_2, \widetilde{\lambda}_1, \widetilde{\lambda}_2$:

Next, it is easy to see that $\sqrt{\lambda_1/\lambda_2}$ increases monotonically, as we increase σ_{sp}. The same is true, for $\widetilde{\lambda}_1/\widetilde{\lambda}_2$, and similarly λ_1/λ_2. Both of these hold since, once again the rate of increase $\frac{\partial \lambda_1}{\partial \sigma_{\text{sp}}} > \frac{\partial \lambda_2}{\partial \sigma_{\text{sp}}}$, and $\frac{\partial \widetilde{\lambda}_1}{\partial \sigma_{\text{sp}}} > \frac{\partial \widetilde{\lambda}_2}{\partial \sigma_{\text{sp}}}$ — both of which are derived from the expressions in Lemma J.4, taking $\sigma_{\text{in}} \ll \gamma$, and $\gamma \ll \sqrt{d_{\text{sp}}}$. Consequently, $\lambda_1\widetilde{\lambda}_1/\lambda_2\widetilde{\lambda}_2$ also increases as we increase σ_{sp}.

Finally, we will focus on the expression $\widetilde{\lambda}_1/\widetilde{\lambda}_2 - \lambda_1/\lambda_2$. Here, we will first see that $\exists \sigma_{\text{sp}_2}$ such that this expression is positive $\forall \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$. If we evaluate the expressions: $\widetilde{\lambda}_1/\widetilde{\lambda}_2$ and λ_1/λ_2, we will note that:

$$\widetilde{\lambda}_1/\widetilde{\lambda}_2 = \frac{1 + \widetilde{z}}{1 - \widetilde{z}} \quad \lambda_1/\lambda_2 = \frac{1 + z}{1 - z}, \tag{184}$$

$$\widetilde{z} := \sqrt{\frac{\gamma^2 d_{\text{sp}} - 4\gamma^2\left(d_{\text{sp}}/2 + \sigma_{\text{sp}}^2/2\right)}{\left(d_{\text{sp}}/2 + \sigma_{\text{sp}}^2/2 + \gamma^2\right)^2} + 1} \tag{185}$$

$$z := \sqrt{\frac{\gamma^2 d_{\text{sp}} - 4\left(\gamma^2(1 + 1/3d_{\text{in}}) + \sigma_{\text{in}}^2(1/3 - 1/3d_{\text{in}})\right)\left(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3 + 1/6\right)}{\left(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3 + 1/6 + \gamma^2(1 + 1/3d_{\text{in}}) + \sigma_{\text{in}}^2(1/3 - 1/3d_{\text{in}})\right)^2} + 1} \tag{186}$$

When $\sigma_{\text{sp}} \ll d_{\text{sp}}$, $\widetilde{z}/z = \mathcal{O}\left(\frac{\sqrt{(d_{\text{sp}}/2)^2 - \gamma^2 d_{\text{sp}} - 2\gamma^2\sigma_{\text{sp}}^2}}{\sqrt{(d_{\text{sp}}/2)^2 - \gamma^2 d_{\text{sp}} - (8/3)\gamma^2\sigma_{\text{sp}}^2}}\right)$. Since $\widetilde{z}/z > 1$ in the region: $0 \leqslant \widetilde{z}, z \leqslant 1$, from the properties of the function $x \mapsto 1+x/1-x$, we can argue that $\widetilde{\lambda}_1/\widetilde{\lambda}_2 > \lambda_1/\lambda_2$. Thus, the term, $\widetilde{\lambda}_1/\widetilde{\lambda}_2 - \lambda_1/\lambda_2$ is positive. Additionally, in the same region, *i.e.*, for some $d_{\text{sp}} \geqslant d_{\text{sp}_0}$, we can argue that $\widetilde{z}$ and z remain constant (up to some approximation terms). Thus, the expression $\widetilde{\lambda}_1/\widetilde{\lambda}_2 - \lambda_1/\lambda_2$ remains stable for small enough σ_{sp}.

Thus, when $\alpha - \beta$ increases, and consequently $\tan(\alpha - \beta)^2$ increases, the denominator term (in $\tan(2\theta)$) decreases monotonically. Recall that numerator also is increasing monotonically under conditions: $\gamma, \sigma_{\text{in}}, \sigma_{\text{sp}} \ll d_{\text{sp}}$, when we increase σ_{sp} from 0 to a positive value. Because of this monotonic behavior there would necessarily exist σ_{sp_1} such that as $\sigma_{\text{sp}} \geqslant \sigma_{\text{sp}_1}$, we have: $\tan(\theta) \geqslant 5\gamma/\tau_0\sqrt{d_{\text{sp}}}$. Similarly, there would exist $\sigma_{\text{sp}_2} \geqslant \sigma_{\text{sp}_1}$, such that $\forall \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$, $\tan(\theta) \leqslant 9\gamma/\tau_0\sqrt{d_{\text{sp}}}$.

Bounded nature of τ:

The expression for τ is simply:

$$\tau = \sqrt{\frac{1+z}{1-z}} \tag{187}$$

$$z = \sqrt{\frac{\gamma^2 d_{\text{sp}} - 4\left(\gamma^2(1+1/3d_{\text{in}}) + \sigma_{\text{in}}^2(1/3 - 1/3d_{\text{in}})\right)\left(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3 + 1/6\right)}{\left(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3 + 1/6 + \gamma^2(1+1/3d_{\text{in}}) + \sigma_{\text{in}}^2(1/3 - 1/3d_{\text{in}})\right)^2} + 1} \tag{188}$$

Since $\gamma, \sigma_{\text{in}} \ll d_{\text{sp}}$, $z = \mathcal{O}\left(\sqrt{1 - \frac{\gamma^2 d_{\text{sp}} - 8/3\gamma^2\sigma_{\text{sp}}^2}{(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3)^2}}\right)$. Further, z increases monotonically, since the term $(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3)^2$ increases at a rate that is much faster than rate at which $8/3\gamma^2\sigma_{\text{sp}}^2$ increases, when $d_{\text{sp}} \geqslant d_{\text{sp}0}$ (or, $\frac{\mathrm{d}g(\sigma_{\text{sp}})}{\mathrm{d}\sigma_{\text{sp}}} > 0$ for large enough d_{sp} where $g(\sigma_{\text{sp}}) = \sqrt{1 - \frac{\gamma^2 d_{\text{sp}} - 8/3\gamma^2\sigma_{\text{sp}}^2}{(d_{\text{sp}}/2 + 2\sigma_{\text{sp}}^2/3)^2}}$. Consequently, τ increases monotonically as σ_{sp} increases. Thus, there would exist some $\sigma'_{\text{sp}2}$ such that $\forall \sigma_{\text{sp}} \leqslant \sigma'_{\text{sp}2}$ we have $\tau \leqslant p\tau_0$ (where $p \in (1,2)$). Now, both τ and $\tan(\theta)$ increase monotonically, but the rate of increase of $\tan(\theta)$ is much faster than τ. Recall, that to in the argument for increase in $\tan(\theta)$, it was sufficient for z to remain constant, *i.e.*, remain close to its value at $\sigma_{\text{sp}} = 0$ for the term $\tan(2\theta)$ to increase. Thus, the condition for $\tau \leqslant p\tau_0$ (for $p \in (1,2)$) is satisfied more easily, and $\sigma'_{\text{sp}2} > \sigma_{\text{sp}2}$.

Note that our arguments above do not necessarily treat d_{sp} as a free parameter. In fact, recall that d_{sp} controls the rate at which $\alpha - \beta$ increases, given by $\mathcal{O}(\gamma/\sqrt{d_{\text{sp}}})$. Hence, $\gamma/\sqrt{d_{\text{sp}}}$ cannot be exactly 0. The key point here is that our required lower bound on τ is $\Omega(1/\tau_0^2)$ and $\tau_0 \simeq \sqrt{d_{\text{sp}}}/\gamma$. Thus the required conditions on c_1/c_3 also relax, and do so with quadratic rates.

Combining arguments on $\tau, \tan(\theta)$, we conclude that, when $\sigma_{\text{in}} \ll \gamma$ and $\gamma \ll d_{\text{sp}}$ (conditions in Theorem H.12), we can show that $\exists \sigma_{\text{sp}1}, \sigma_{\text{sp}2}$ such that for the range of $\sigma_{\text{sp}1} \leqslant \sigma_{\text{sp}} \leqslant \sigma_{\text{sp}2}$, $5\gamma/\tau_0\sqrt{d_{\text{sp}}} \leqslant \tan\theta \leqslant 9\gamma/\tau_0\sqrt{d_{\text{sp}}}$. Further, there exists $p \in (1,2)$ such that $\tau \leqslant p\tau_0$. □

Lemma J.6. *As we increase σ_{sp}, the value of $\cot(\alpha)$ decreases monotonically,* i.e. $\cot(\alpha_0) \geqslant \cot\alpha,\ \forall \sigma_{\text{sp}}$. *Furthermore, when $\gamma, \sigma_{\text{in}} \ll \sqrt{d_{\text{sp}}}$ (conditions from Theorem H.12), we get $(1+\epsilon_0)\frac{\sqrt{d_{\text{sp}}}}{\gamma} \geqslant \tan\alpha_0 \geqslant (1-\epsilon_0)\frac{\sqrt{d_{\text{sp}}}}{\gamma}$ for some small* $1 > \epsilon_0 > 0$.

Proof. Let us look at the expression of $\tan(\alpha)$ from J.4.

$$\tan(\alpha) = \frac{1}{\gamma\sqrt{d_{\text{sp}}}}\left(\frac{d_{\text{sp}}}{2} + \frac{2\sigma_{\text{sp}}^2}{3} + \frac{1}{6} - \left(\gamma^2\left(1 + \frac{1}{3d_{\text{in}}}\right) + \frac{\sigma_{\text{in}}^2}{3}\left(1 - \frac{1}{d_{\text{in}}}\right)\right)\right.$$
$$\left. + \sqrt{\gamma^2 d_{\text{sp}} + \left(\left(\gamma^2\left(1 + \frac{1}{3d_{\text{in}}}\right) + \frac{\sigma_{\text{in}}^2}{3}\left(1 - \frac{1}{d_{\text{in}}}\right)\right) - \left(\frac{d_{\text{sp}}}{2} + \frac{2\sigma_{\text{sp}}^2}{3} + \frac{1}{6}\right)\right)^2}\right)$$

As we increase σ_{sp}, the term $2\sigma_{\text{sp}}^2/3$ monotonically increases in the numerator. Also, the term inside the $\sqrt{\cdot}$ expression: $\left(\left(\gamma^2\left(1 + \frac{1}{3d_{\text{in}}}\right) + \frac{\sigma_{\text{in}}^2}{3}\left(1 - \frac{1}{d_{\text{in}}}\right)\right) - \left(\frac{d_{\text{sp}}}{2} + \frac{2\sigma_{\text{sp}}^2}{3} + \frac{1}{6}\right)\right)$ monotonically increases in magnitude. Thus, it is evident that $\tan(\alpha)$ would monotonically increase as we increase σ_{sp}, and consequently $\cot(\alpha)$ would decrease with increase in σ_{sp}, making $\cot(\alpha_0)$ the maximum value of $\cot\alpha$ for fixed $\gamma, d_{\text{sp}}, \sigma_{\text{in}}, d_{\text{in}}$.

Next, we look at the value of $\tan\alpha_0$ under the condition $\gamma, \sigma_{\text{in}} \ll \sqrt{d_{\text{sp}}}$. Here we see that,

$$\tan(\alpha_0) = \frac{\frac{d_{\text{sp}}}{2} + \mathcal{O}(\gamma^2 + \sigma_{\text{in}}^2) + \gamma\sqrt{d_{\text{sp}}}\left(\sqrt{d_{\text{sp}}}/2\gamma + \mathcal{O}(1/\sqrt{d_{\text{sp}}}\gamma)\right)}{\gamma\sqrt{d_{\text{sp}}}} = \frac{\sqrt{d_{\text{sp}}}}{\gamma} + \Theta(1/\sqrt{d_{\text{sp}}}) \tag{189}$$

Thus, when d_{sp} is sufficiently large, compared to $\sigma_{\text{in}}, \gamma$, $(1+\epsilon_0)\frac{\sqrt{d_{\text{sp}}}}{\gamma} \geqslant \tan\alpha_0 \geqslant (1-\epsilon_0)\frac{\sqrt{d_{\text{sp}}}}{\gamma}$, for some small $1 > \epsilon_0 > 0$. □

Lemma J.7. *As we increase σ_{sp}, the value of τ increases monotonically,* i.e. $\tau_0 \leqslant \tau,\ \forall \sigma_{\text{sp}}$. *Additionally, when $\gamma, \sigma_{\text{in}} \ll d_{\text{sp}}$ (conditions from Theorem H.12), we have $(1+\epsilon_1)\frac{\sqrt{d_{\text{sp}}}}{\gamma} \geqslant \tau_0 \geqslant (1-\epsilon_1)\frac{\sqrt{d_{\text{sp}}}}{\gamma}$ for some small $1 > \epsilon_1 > 0$.*

Proof. The proof of this lemma follows from arguments made in Lemma J.6 and Lemma J.5. Recall that:

$$\tau_0 = \sqrt{\frac{1+z_0}{1-z_0}} \tag{190}$$

$$\tau = \sqrt{\frac{1+z}{1-z}} \tag{191}$$

$$z = \sqrt{\frac{\gamma^2 d_{\text{sp}} - 4\left(\gamma^2(1+{}^1\!/_{3d_{\text{in}}}) + \sigma_{\text{in}}^2({}^1\!/_3 - {}^1\!/_{3d_{\text{in}}})\right)\left(d_{\text{sp}}/2 + {}^{2\sigma_{\text{sp}}^2}\!/_3 + 1/6\right)}{\left(d_{\text{sp}}/2 + {}^{2\sigma_{\text{sp}}^2}\!/_3 + 1/6 + \gamma^2(1+{}^1\!/_{3d_{\text{in}}}) + \sigma_{\text{in}}^2({}^1\!/_3 - {}^1\!/_{3d_{\text{in}}})\right)^2} + 1}, \tag{192}$$

where z_0 is the value that z takes at $\sigma_{\text{sp}} = 0$. In the second part of Lemma J.5 we have already argued that τ increases monotonically as σ_{sp} increases from $0 \to \sigma_{\text{sp}_2}$. Thus, now we are only left to reason about τ_0. We can see that z_0 evaluates to:

$$z_0 = \sqrt{\frac{\gamma^2 d_{\text{sp}} - 4\left(\gamma^2(1+{}^1\!/_{3d_{\text{in}}}) + \sigma_{\text{in}}^2({}^1\!/_3 - {}^1\!/_{3d_{\text{in}}})\right)\left(d_{\text{sp}}/2 + 1/6\right)}{\left(d_{\text{sp}}/2 + 1/6 + \gamma^2(1+{}^1\!/_{3d_{\text{in}}}) + \sigma_{\text{in}}^2({}^1\!/_3 - {}^1\!/_{3d_{\text{in}}})\right)^2} + 1} \tag{193}$$

Under conditions of Theorem H.12, we know $\sigma_{\text{in}} \ll \gamma \ll \sqrt{d_{\text{sp}}}$. Taking $\sigma_{\text{in}} \ll \gamma$, we get $z_0 \simeq \sqrt{1 - {}^{4\gamma^2}\!/_{d_{\text{sp}}}}$. Now taking $\gamma \ll \sqrt{d_{\text{sp}}}$ we can use Taylor approximation to approximate $\sqrt{1-x^2}$ with $1 - {}^{x^2}\!/_2$ when x is close to 0. Consequently, we get $z_0 \simeq 1 - {}^{2\gamma^2}\!/_{d_{\text{sp}}}$. Plugging this in to $\sqrt{1+z_0}/\sqrt{1-z_0}$ we get $\tau_0 \simeq \sqrt{{}^{d_{\text{sp}}}\!/_\gamma}$. Thus, we can conclude that $\exists \epsilon_1 \approx 0$ such that $(1+\epsilon_1)\frac{\sqrt{d_{\text{sp}}}}{\gamma} \geqslant \tau_0 \geqslant (1-\epsilon_1)\frac{\sqrt{d_{\text{sp}}}}{\gamma}$. □

Lemma J.8. *Under conditions on $\gamma, d_{\text{sp}}, \sigma_{\text{in}}$ in Theorem H.12, and bounded range of $\sigma_{\text{sp}_1}\sigma_{\text{sp}} \leqslant \sigma_{\text{sp}_2}$ (from Lemma J.5), we can show the following is true: $c_1, c_3, c_4 > 0$ and $c_2 < 0$.*

Proof. By definition, $c_1, c_4 \geqslant 0$. In the proof on the lower bound over c_1/c_3, we argue that under conditions on problem parameters defined in Theorem H.12 and for a bounded range of noise in target (Lemma J.5), c_1/c_3 remains positive. Hence, $c_3 > 0$. Now, consider the expression for $c_2 = -1 + \frac{\cot(\alpha)\tan(\theta)}{\tau}$.

Primarily, we note from Lemma J.6 and Lemma J.5 that:

$$\frac{\cot(\alpha)\tan(\theta)}{\tau} \leqslant \frac{\cot(\alpha_0)^{9\gamma}\!/\!\sqrt{d_{\text{sp}}}\,\tau_0}{\tau_0} = \mathcal{O}({}^{\gamma^4}\!/_{d_{\text{sp}}{}^2}),$$

since $\cot\alpha_0 \simeq {}^\gamma\!/\!\sqrt{d_{\text{sp}}}$ and $\tau_0 \simeq \sqrt{{}^{d_{\text{sp}}}\!/_\gamma}$.. As a result, we can conclude $c_2 < 0$. □

Lemma J.9 (0-1 error of a classifier on target)**.** *Assume a classifier of the form $w = l_1 \cdot w_{\text{in}} + l_2 \cdot w_{\text{sp}}$ where $l_1, l_2 \in \mathbb{R}$ and $w_{\text{in}} = [w^\star, 0, ..., 0]^\top$, and $w_{\text{sp}} = [0, ..., 0, \mathbf{1}_{d_{\text{sp}}}/\sqrt{d_{\text{sp}}}]^\top$. Then the target accuracy of this classifier is given by $0.5 \cdot \text{erfc}\left(-\frac{l_1 \cdot \gamma}{\sqrt{2} \cdot l_2 \cdot \sigma_{\text{sp}}}\right)$.*

Proof. Assume $(x, y) \sim \mathrm{P_T}$. Accuracy of w is given by $\mathbb{E}_{\mathrm{P_T}}\left[(\mathrm{sign}\left(w^\top x\right) = y)\right]$.

$$\begin{aligned}
\mathbb{E}_{\mathrm{P_T}}\left[\mathrm{sign}\left(w^\top x\right) = y\right] &= \mathbb{E}_{\mathrm{P_T}}\left[y \cdot \mathrm{sign}\left(w^\top x\right) = 1\right] \\
&= \mathbb{E}_{\mathrm{P_T}}\left[y \cdot (w^\top x) > 0\right] \\
&= \mathbb{E}_{\mathrm{P_T}}\left[y \cdot (x^\top (l_1 \cdot w_{\mathrm{in}} + l_2 \cdot w_{\mathrm{sp}})) > 0\right] \\
&= \mathbb{E}_{\mathrm{P_T}}\left[y \cdot (\gamma \cdot l_1 \cdot y + l_2 \cdot \sigma_{\mathrm{sp}}) > 0\right] \\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[(\gamma \cdot l_1 + y \cdot l_2 \cdot \sigma_{\mathrm{sp}} \cdot z) > 0\right] \\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[y \cdot l_2 \cdot \sigma_{\mathrm{sp}} \cdot z > -\gamma \cdot l_1\right] \\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[l_2 \cdot \sigma_{\mathrm{sp}} \cdot z > -\gamma \cdot l_1\right] \\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[z > -\frac{\gamma \cdot l_1}{l_2 \cdot \sigma_{\mathrm{sp}}}\right]
\end{aligned}$$

Using the definition of erfc function, we get the aforementioned accuracy expression. □

Lemma J.10. *For $\sigma > 0$ and $\mu \in \mathbb{R}$, we have*

$$\begin{aligned}
g(\mu, \sigma) &:= \mathbb{E}_{z \sim \mathcal{N}(0,\sigma)}\left[\exp\left(-\left|\mu + z\right|\right)\right] && (194) \\
&= \frac{1}{2}\left(\exp\left(\sigma^2/2 - \mu\right) \cdot \mathrm{erfc}\left(-\mu/\sqrt{2}\sigma + \sigma/\sqrt{2}\right) + \exp\left(\sigma^2/2 + \mu\right) \cdot \mathrm{erfc}\left(\mu/\sqrt{2}\sigma + \sigma/\sqrt{2}\right)\right) && (195)
\end{aligned}$$

Proof. The proof uses simple algebra and the definition of erfc function.

$$\begin{aligned}
g(\mu, \sigma) &:= \mathbb{E}_{z \sim \mathcal{N}(0,\sigma)}\left[\exp\left(-\left|\mu + z\right|\right)\right] \\
&= \frac{1}{\sqrt{2\pi}} \int_z \exp\left(-\left|\mu + z\right|\right) \cdot \exp\left(-\frac{z^2}{2\sigma^2}\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\left|\mu + z\right|\right) \cdot \exp\left(-\frac{z^2}{2\sigma^2}\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\mu}^{\infty} \exp\left(-\mu + z\right) \cdot \exp\left(-\frac{z^2}{2\sigma^2}\right) dz + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\mu} \exp\left(\mu + z\right) \cdot \exp\left(-\frac{z^2}{2\sigma^2}\right) dz \\
&= \exp\left(\sigma^2/2 - \mu\right) \int_{\frac{-\mu}{\sqrt{2}\sigma} + \frac{\sqrt{2}\sigma}{2}}^{\infty} \exp(-z^2) dz + \exp\left(\sigma^2/2 + \mu\right) \int_{-\infty}^{\frac{-\mu}{\sqrt{2}\sigma} - \frac{\sqrt{2}\sigma}{2}} \exp(-z^2) dz \\
&= \frac{1}{2}\left(\exp\left(\sigma^2/2 - \mu\right) \cdot \mathrm{erfc}\left(-\mu/\sqrt{2}\sigma + \sigma/\sqrt{2}\right) + \exp\left(\sigma^2/2 + \mu\right) \cdot \mathrm{erfc}\left(\mu/\sqrt{2}\sigma + \sigma/\sqrt{2}\right)\right)
\end{aligned}$$

□

Lemma J.11. *For $\mu_t \geqslant \mu_0$ and $|\sigma_t| \leqslant \sigma_0$, we have for all t:*

$$\frac{A_3 - A_1\sigma_t - A_2\sigma_t}{A_1 - A_2} \leqslant p(\sigma_0, \mu_0)\,,$$

where A_1, A_2 and A_3 defined in (127), (128)*, and* (129)*, and p is defined in* (208).

Proof. Recall the definition of A_1, A_2, and A_3.

$$A_1(\mu_t, \sigma_t) = \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \mathrm{erfc}\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\,, \quad (196)$$

$$A_2(\mu_t, \sigma_t) = \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \mathrm{erfc}\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)\,, \quad (197)$$

$$A_3(\mu_t, \sigma_t) = \frac{2\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{\mu_t^2}{2\sigma_t^2}\right)\,. \quad (198)$$

We now use upper bounds and lower bounds on erfc as in Lemma J.1. In particular, we have the following bounds on A_1 and A_2:

$$A_1 \leqslant \frac{2}{\sqrt{\pi}} \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \frac{\exp\left(-\frac{\sigma_t^2}{2} + \mu_t - \mu_t^2/(2\cdot\sigma_t^2)\right)}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 4/\pi}} \tag{199}$$

$$= \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 4/\pi}} . \tag{200}$$

$$A_1 \geqslant \frac{2}{\sqrt{\pi}} \exp\left(\frac{\sigma_t^2}{2} - \mu_t\right) \cdot \frac{\exp\left(-\frac{\sigma_t^2}{2} + \mu_t - \mu_t^2/(2\cdot\sigma_t^2)\right)}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} \tag{201}$$

$$= \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(-\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} . \tag{202}$$

$$A_2 \leqslant \frac{2}{\sqrt{\pi}} \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \frac{\exp\left(-\frac{\sigma_t^2}{2} - \mu_t - \mu_t^2/(2\cdot\sigma_t^2)\right)}{\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 4/\pi}} \tag{203}$$

$$= \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 4/\pi}} . \tag{204}$$

$$A_2 \geqslant \frac{2}{\sqrt{\pi}} \exp\left(\frac{\sigma_t^2}{2} + \mu_t\right) \cdot \frac{\exp\left(-\frac{\sigma_t^2}{2} - \mu_t - \mu_t^2/(2\cdot\sigma_t^2)\right)}{\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} \tag{205}$$

$$= \frac{2}{\sqrt{\pi}} \frac{\exp\left(-\mu_t^2/(2\cdot\sigma_t^2)\right)}{\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}} . \tag{206}$$

Using these bounds, we get:

$$\frac{A_3 - A_1\sigma_t - A_2\sigma_t}{A_1 - A_2} \leqslant p(\sigma_t, \mu_t) , \tag{207}$$

where

$$p(\sigma_t, \mu_t) = \frac{\sqrt{2} \cdot \left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}\right) \left(\frac{-\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}} + \sqrt{\left(\frac{-\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}\right)}{\frac{\sqrt{2}\mu_t}{\sigma_t} + \sqrt{\left(\frac{\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 4/\pi} - \left(\sqrt{\left(\frac{-\mu_t}{\sqrt{2}\sigma_t} + \frac{\sigma_t}{\sqrt{2}}\right)^2 + 2}\right)} . \tag{208}$$

We observe that the RHS of (208) increases with σ_t and decreases with μ_t and takes the maximum value at boundary points σ_0 and μ_0.

□